U.S. DEPARTMENT OF COMMERCE
National Technical Information Service

PB-285 436

Population at Risk to Various Air Pollution Exposures Data Base 'POPATRISK'

System Sciences, Inc, Chapel Hill, N C

Prepared for

Health Effects Research Lab, Research Triangle Park, N C

Jun 78

United States
Environmental Protection
Agency

Health Effects Research
Laboratory
Research Triangle Park NC 27711

EPA-600/1-78-051
June 1978

Research and Development

PB 285 436

EPA

Population at Risk to Various Air Pollution Exposures : Data Base "Popatrisk"

REPRODUCED BY
NATIONAL TECHNICAL
INFORMATION SERVICE
U. S. DEPARTMENT OF COMMERCE
SPRINGFIELD, VA. 22161

RESEARCH REPORTING SERIES

Research reports of the Office of Research and Development, U.S. Environmental Protection Agency, have been grouped into nine series. These nine broad categories were established to facilitate further development and application of environmental technology. Elimination of traditional grouping was consciously planned to foster technology transfer and a maximum interface in related fields. The nine series are:

1. Environmental Health Effects Research
2. Environmental Protection Technology
3. Ecological Research
4. Environmental Monitoring
5. Socioeconomic Environmental Studies
6. Scientific and Technical Assessment Reports (STAR)
7. Interagency Energy-Environment Research and Development
8. "Special" Reports
9. Miscellaneous Reports

This report has been assigned to the ENVIRONMENTAL HEALTH EFFECTS RESEARCH series. This series describes projects and studies relating to the tolerances of man for unhealthful substances or conditions. This work is generally assessed from a medical viewpoint, including physiological or psychological studies. In addition to toxicology and other medical specialities, study areas include biomedical instrumentation and health research techniques utilizing animals — but always with intended application to human health measures.

This document is available to the public through the National Technical Information Service, Springfield, Virginia 22161.

TECHNICAL REPORT DATA

(Please read Instructions on the reverse before completing)

1. REPORT NO. EPA-600/1-78-051	2.	3. RECIPIENT'S ACCESSION NO. PB285436
4. TITLE AND SUBTITLE Population at Risk to Various Air Pollution Exposures: Data Base "Popatrisk"		5. REPORT DATE June 1978
		6. PERFORMING ORGANIZATION CODE
7. AUTHOR(S) Sandor J. Freedman Elsa Lewis-Heise Joseph D. Wilson Albert V. Hardy		8. PERFORMING ORGANIZATION REPORT NO.
9. PERFORMING ORGANIZATION NAME AND ADDRESS System Sciences, Inc. P. O. Box 2345 Chapel Hill, North Carolina 27514		10. PROGRAM ELEMENT NO. 1AA601
		11. CONTRACT/GRANT NO. Contract No. 68-02-2269
12. SPONSORING AGENCY NAME AND ADDRESS Health Effects Research Laboratory RTP,NC Office of Research and Development U.S. Environmental Protection Agency Research Triangle Park, North Carolina 27711		13. TYPE OF REPORT AND PERIOD COVERED Final report covering Oct.1975 - Dec. 1977
		14. SPONSORING AGENCY CODE EPA 600/11
15. SUPPLEMENTARY NOTES		

16. ABSTRACT

The work reported herein was undertaken to provide the EPA with a user-oriented data base containing recent county-based information, for all counties in the contiguous United States, on population demographics, population mobility, climatology, emissions, air quality, and age-adjusted death rates.

The completed data base, called "POPATRISK," contains approximately 27.5 million characters and is in SYSTEM 2000, Version 2.80 format, facilitating access with minimal user computer training. Population demographics are as of the 1970 Census; population mobility is described spanning the years 1965 to 1970 for 6 sex-race categories in 7 age groupings for both "in" and "out" migrants" climatology information contains county summaries of temperature, precipitation and hours of sunshine; county point and area source emission estimates are provided for 5 criteria pollutants--TSP, SO_2, NO_2, CO, and Ozone--based on the NEDS-USER file; air quality information is based on 1974 data contained in SAROAD; age-adjusted death rates were computed for the combined years 1969, 1970, and 1971 for 4 sex-race categories in 50 groupings of ICDA categories (8th revision).

17. KEY WORDS AND DOCUMENT ANALYSIS

a. DESCRIPTORS	b. IDENTIFIERS/OPEN ENDED TERMS	c. COSATI Field/Group
data file air quality demography population mortality	Population at risk	05A 09B 06F

18. DISTRIBUTION STATEMENT Release to Public	19. SECURITY CLASS *(This Report)* Unclassified	21. NO. OF PAGES 144
	20. SECURITY CLASS *(This page)* Unclassified	22. PRICE A07-A01

EPA Form 2220-1 (9-73)

EPA-600/1-78-051
June 1978

POPULATION AT RISK TO VARIOUS AIR POLLUTION EXPOSURES: DATA BASE "POPATRISK"

by

Sandor J. Freedman
Joseph D. Wilson
Elsa Lewis-Heise
Albert V. Hardy

System Sciences, Inc.
Chapel Hill, North Carolina 27514

Contract No. 68-02-2269

Project Officer

William C. Nelson
Statistics and Data Management Office
Health Effects Research Laboratory
Research Triangle Park, North Carolina 27711

Health Effects Research Laboratory
Office of Research and Development
U.S. Environmental Protection Agency
Research Triangle Park, North Carolina 27711

i.a

DISCLAIMER

This report has been reviewed by the Health Effects Research Laboratory, U.S. Environmental Protection Agency, and approved for publication. Approval does not signify that the contents necessarily reflect the views and policies of the U.S. Environmental Protection Agency, nor does mention of trade names or commercial products constitute endorsement or recommendation for use.

FOREWORD

The many benefits of our modern, developing, industrial society are accompanied by certain hazards. Careful assessment of the relative risk of existing and new man-made environmental hazards is necessary for the establishment of sound regulatory policy. These regulations serve to enhance the quality of our environment in order to promote the public health and welfare and the productive capacity of our Nation's population.

The Health Effects Research Laboratory, Research Triangle Park, conducts a coordinated environmental health research program in toxicology, epidemiology, and clinical studies using human volunteer subjects. These studies address problems in air pollution, non-ionizing radiation, environmental carcinogenesis and the toxicology of pesticides as well as other chemical pollutants. The Laboratory participates in the development and revision of air quality criteria documents on pollutants for which national ambient air quality standards exist or are proposed, provides the data for registration of new pesticides or proposed suspension of those already in use, conducts research on hazardous and toxic materials, and is primarily responsible for providing the health basis for non-ionizing radiation standards. Direct support to the regulatory function of the Agency is provided in the form of expert testimony and preparation of affidavits as well as expert advice to the Administrator to assure the adequacy of health care and surveillance of persons having suffered imminent and substantial endangerment of their health.

The data base described in this report has been developed to provide the capability to examine easily and quickly the available county level information on air quality; population including social-economic, demographic, and migration factors, climatology, and mortality. The original information has been collected by various government agencies but has not previously been combined into a single data file. Use of this file will permit more accurate estimates of the populations exposed to various air pollutant levels and a better assessment of the geographic variability between air quality and mortality.

F. G. Hueter, Ph. D.
Acting Director,
Health Effects Research Laboratory

ABSTRACT

The work reported herein was undertaken to provide the EPA with a user-oriented data base containing recent county-based information, for all counties in the contiguous United States, on population demographics, population mobility, climatology, emissions, air quality, and age-adjusted death rates.

The completed data base, called "POPATRISK," contains approximately 27.5 million characters and is in SYSTEM 2000, Version 2.80 format, facilitating access with minimal user computer training. Population demographics are as of the 1970 Census; population mobility is described spanning the years 1965 to 1970 for 6 sex-race categories in 7 age groupings for both "in" and "out" migrants; climatology information contains county summaries of temperature, precipitation and hours of sunshine; county point and area source emission estimates are provided for 5 criteria pollutants--TSP, SO_2, NO_2, CO, and Ozone--based on the NEDS-USER file; air quality information is based on 1974 data contained in SAROAD; age-adjusted death rates were computed for the combined years 1969, 1970, and 1971 for 4 sex-race categories in 50 groupings of ICDA categories (8th revision).

Sample applications of the data base are provided herein. A detailed manual documenting the county identification codes to be used in retrieving data has been provided under separate cover. Also included in the geocoding manual is a cross-reference table showing the relationships among POPATRISK and the SAROAD, FIPS, Census and NCHS geocoding schemes.

This report was submitted in fulfillment of Contract No. 68-02-2269 by Systems Sciences, Inc. under the sponsorship of the U.S. Environmental Protection Agency. This report covers a period from October 29, 1975, to December 31, 1977, and work was completed as of December 31, 1977.

CONTENTS

Abstract ii
Figures iv
Tables v
Acknowledgements vii

1. Introduction 1
2. POPATRISK Data Base Features 3
    - System 2000 Version 2.80 Features 3
    - POPATRISK Data Base Structure 4
    - Components and Sources of Data 9
    - Variations in Geocoding Schemes Among Sources 24
3. Use of Data Base "POPATRISK" 26
4. Sample Retrievals Using POPATRISK 29
5. Data Base Development Procedures 44
    - Considerations Leading to Choice of Data Items 44
    - History of Development 46
    - Methodologies for Processing and Loading 47
    - Use of Removable Disk Pack and Archiving Procedures 66

Appendix A: Resolution of Geocoding Discrepancies for Massachusetts NEDS and SAROAD Data 69
Appendix B: Documentation of Data Tapes and Runstreams 76
Appendix C: Other Studies 88

FIGURES

Number		Page
1	POPATRISK heirarchical structure	8
2	Overview of POPATRISK data sources and development procedures, Phase 1	51
3	Overview of POPATRISK data sources and development procedures, Phase 2	58
C-1	1973 Monitoring Station Sample Distribution	92
C-2	Fit of present model with all valid 1973 SAROAD population based monitoring	115
C-3	"Inherent" variability of SAROAD data for intracounty population based monitoring	116
C-4	Comparison of fitted normal curve for predicted air quality and "inherent" variability	117
C-5	Variation of R^2 with data set size	118

TABLES

Number		Page
1	POPATRISK Data Base Definition in System 2000 Format	5
2	POPATRISK Data Base Component Description	10
3	Aggregations of Mortality Categories for POPATRISK Age-Adjusted Deaths	21
4	Cluster Descriptions	23
5	Illustrative Retrieval of NEDS Emissions Data for Massachusetts in tons/year	30
6	Illustrative Retrieval of SAROAD Site and Monitoring Data for Durham County, NC	32
7	Illustrative Retrieval of Counties with Highest Calculated Mean SO_2 Monitoring ($\mu g/m^3$)	35
8	Illustrative Retrieval of Counties with Highest Age-Adjusted Auto Fatalities for White Males	37
9	Illustrative Retrieval of Counties with Highest White Male Death Rate from Bronchitis	40
10	Illustrative Retrieval of Counties with Highest White Female Age-Adjusted Death Rate from Breast Cancer	42
11	Illustrative Retrieval of Counties with Highest White Death Rate from Malignant Neoplasm Cluster 1	43
12	Unallocated In-Migration Totals for Selected Counties	48
13	State Unallocated In-Migration Totals for All Races	49
14	Assignment of Massachusetts SAROAD Monitoring to Census Counties	53
15	Independent Cities Recognized and Coded in Data Base	55
16	Assignment of Virginia Independent Cities to POPATRISK Counties	56
A-1	NEDS Points Apportioned into Census Counties by using "Dummy" UTM Coordinates	72
A-2	NEDS Points Apportioned into Census Counties by Tracing Their Name and Address	73
A-3	Total Emissions Summed by Type and by County	74
A-4	Massachusetts Area Emissions Totals by POPATRISK Counties	75

TABLES (continued)

Number		Page
C-1	Listing of Monitoring Methods Included In and Excluded from Survey	90
C-2	Percentage of States' Population in Counties with One or More Monitoring Stations	99
C-3	Cumulative Frequency Tables for Valid 1973 SAROAD Population Based Monitoring of TSP	110
C-4	Maximum Correlation Values, R, for a Sampling of Different Models Investigated	111
C-5	Sample Comparison between Maximum Correlation Coefficients R and Those Obtained from Model	112
C-6	Sample Comparison between Chosen Model and Other "Simple" Models	113
C-7	Effect on Regression Model of Removing 5 "Outlier" Points from 58 Point Data Set	114
C-8-a	County Emission Data and Land Use Characteristics for North Carolina	125
C-8-b	Descriptive Statistics for Model	126
C-9	Vegetation Surface Area Weighting Factors	127
C-10	Spearman Correlations: Annual Average SO_2 Concentrations	128
C-11	Multiple Regression Models: Annual Average SO_2 Concentrations	129

ACKNOWLEDGEMENTS

The continuing support and technical direction of the EPA Project Officer, Dr. William C. Nelson, are gratefully acknowledged. Thanks are due also to Dr. Victor A. Hasselblad of HERL for many helpful conversations. John Van Bruggen of HERL provided information on the ICDA aggregations in current use at HERL.

Paul Cornely, M.D., Ph.D., of System Sciences, Inc., was involved in the interpretation of the cluster analysis and was most helpful also in interpreting some fine points of ICDA coding. Christopher Gordon of System Sciences, Inc. assisted in developing the geocoding scheme for the independent cities in Virginia and also in resolution of other geocoding problems.

Dr. I-Li Huang and Ms. Susan Alston, formerly of System Sciences, Inc., were responsible for some work on the data base in the intial phases of this contract.

The authors are pleased to acknowledge also the assistance of two consultants. Dr. Richard Kopec of the Geography Department, University of North Carolina at Chapel Hill, was responsible for the reallocation of NEDS point emission sources to the Massachusetts FIPS counties. Mr. John Richards of PEDCo Environmental Specialists, Chapel Hill, North Carolina, carried out the study on SO_2 - ground deposition parameters.

Finally, thanks are due to Ms. Signe Wetrogen of the Population Division, Bureau of Census, for her continued interest in the in-out migration data and for providing a special compilation of New York City migration data.

SECTION 1

INTRODUCTION

Compilation of primary source material of nationwide data on population demography and mobility, climatology, emissions of pollutants, air quality and deaths by ICDA cause is the responsibility of a number of different government agencies. The data has heretofore been available mostly in specific format which tends to be unique for each data type. The investigator of potential relationships among those variables has, therefore, often been required to reformat, condense, aggregate and/or disaggregate data which may have conflicting schemes of identifying geographical areas, population age groups and the like.

The main thrust of the work conducted under this project was to provide the EPA with a coherent source of recent data for these variables in which the user, with a minimum of computer training, can retrieve information pertinent to his investigations with maximum speed and flexibility. The user can define the format within the wide capabilities of SYSTEM 2000, Version 2.80.

All compilations of data require some choice of what items should be included and this work is no exception. Considerations leading to this choice were guided by the desirabilities of time, compatibility among data sets, reasonableness in data base size and expected retrieval costs, and anticipated usefulness to the EPA as defined through many discussions between the contractor staff and the EPA project officer.

The sections to follow provide a detailed description of the data base structure and historical development. In particular, Section 2 documents the unique features of the POPATRISK data base including a discussion of the meaning, units, and primary sources of each individual data item. Geocoding unique to this data base, which was necessitated by the variance in geocoding schemes used by the different primary sources, is summarized in this report and detailed under separate cover in the report "POPATRISK Geocoding Manual." Section 3 gives the reader an overview of techniques for performing retrievals although the discussion is by no means meant to be complete. Reference is made to the MRI SYSTEM 2000 Reference Manual for more detailed discussion of the use of a SYSTEM 2000, Version 2.80 data base. Section 4 illustrates some sample results using POPATRISK and is a brief indication of potential uses of the data base. Development procedures used in this work are documented in Section 5. Appendix A contains a discussion of methodologies used to resolve geocoding discrepancies for Massachusetts. Appendix B contains detailed documentation on tapes and runstreams provided EPA with this report. Appendix C describes other studies performed in this work. In particular, a mapped survey of monitoring activity for the criteria pollutants in 1973 and a report of studies correlating ambient TSP levels with county emission estimates are presented.

SECTION 2

POPATRISK DATA BASE FEATURES

SYSTEM 2000, VERSION 2.80 FEATURES

The POPATRISK data base design and contents are a result of analysis and evaluation of available data sources and their potential usefulness to investigators in population-at-risk studies. The SYSTEM 2000 data base management system was selected as the most versatile and cost-effective medium for storage and retrieval of POPATRISK data.

SYSTEM 2000 is a general-purpose data base management system which operates on the UNIVAC 1100 series computers as well as IBM and CDC computers. It provides the user with a comprehensive set of data base management capabilities for developing and utilizing information systems tailored to the specific requirements of the user.

SYSTEM 2000, Version 2.80 offers a full range of capabilities for modification of a data base definition, data base update, and quick-response retrievals. Specifically, a SYSTEM 2000 data base design is flexible, in that additions can be made to the data base definition (structure) at any time and many changes to the existing structure are easily accomplished. The Immediate Access feature offers an English-like user-oriented language which can be used with minimal computer training to update or retrieve data in the data base. Along with the Report Writer feature, Immediate Access provides the user capability to retrieve and print data quickly and in the desired format. Update features enable the user to ADD, CHANGE, or REMOVE data at any level from the entire logical entry down to the individual component level.

A further extension of SYSTEM 2000 retrieval and update features which was used extensively in development of POPATRISK is the Procedural Language Interface feature. It enables the user to manipulate data in a data base through interfacing with COBOL, FORTRAN, or assembly language programs. This feature is particularly useful both in processing and loading large amounts of data and in performing complicated retrievals from logically unrelated sets of data.

A detailed discussion of the SYSTEM 2000 data base management system, its capabilities and command usage may be found in the SYSTEM 2000 Reference Manual, the Version 2.80 Newsletter, and Procedural Language Interface (UM-2).

POPATRISK DATA BASE STRUCTURE

Table 1 shows the POPATRISK data base definition in SYSTEM 2000 format. As an aid to understanding the data base structure, Figure 1 is provided to illustrate the hierarchical relationships of the data within the data base. It is important to note that the data base is described in a county based format since the smallest geographical entity defined in this work is the county. It was decided to adopt a county coding convention adhering as closely as possible to the current list of NADB SAROAD codes to provide a system most adaptable to the data base and to the user. Some differences do exist, however, between the POPATRISK geocoding scheme and the SAROAD scheme. These are documented below. Data in the data base are stored in groupings referred to as logical entries. One logical entry contains all information pertaining to a particular county, thus there is one logical entry of data for each county represented in the data base. Within a logical entry, data are organized in tree structure into repeating groups of either related or disjoint data sets. Each repeating group represents a different type of data and components within the group define in detail the data sets contained therein.

In Figure 1, each level 1 repeating group is a descendant of the Level 0 or Entry Level data and the Level 2 repeating group is a descendant of a Level 1 repeating group. All data sets connected by a line follow a branch of the tree structure and are considered related data sets, however, they are disjoint to data sets connected to different lines (branches) from the Entry Level. For example, MONITORING DATA repeating group C300 is a

TABLE 1. POPATRISK DATA BASE DEFINITION IN SYSTEM 2000 FORMAT

```
SYSTEM RELEASE NUMBER   2.80B
DATA BASE NAME IS POPATRISK
DEFINITION NUMBER 11
DATA BASE CYCLE   25470
      1*  STATE ID (NAME XX)
      2*  COUNTY ID (NAME XXXX)
      3*  ST-COUNTY ID (NAME X(6))
     10*  TOTAL POPULATION (INTEGER NUMBER 9(8))
     11*  PERCENT NONWHITE (DECIMAL NUMBER 99.9)
     12*  PERCENT BLACK (DECIMAL NUMBER 99.9)
     13*  PERCENT OVER 64 (DECIMAL NUMBER 99.9)
     14*  PERCENT FEMALE (DECIMAL NUMBER 99.9)
     15*  POP PER SQ MI (DECIMAL NUMBER 9(5).9)
     16*  PERCENT URBAN (DECIMAL NUMBER 99.9)
     17*  % EMP OTHER CNTY (NON-KEY  DECIMAL NUMBER 99.9)
     18*  INCOME >10K/<5K (NON-KEY  DECIMAL NUMBER 99.99)
     19*  INCOME >15K/<10K (NON-KEY  DECIMAL NUMBER 99.99)
     20*  INCOME >15K/<5K (NON-KEY  DECIMAL NUMBER 99.99)
     21*  % USE PUB TRANS (NON-KEY  DECIMAL NUMBER 99.9)
     22*  ALCOHOL SALE >18 (DECIMAL NUMBER 999.9)
     23*  GAS SALES/CAPITA (DECIMAL NUMBER 999.9)
     30*  MEAN TEMP JAN (NON-KEY  DECIMAL NUMBER 99.9)
     31*  MEAN TEMP JUL (NON-KEY  DECIMAL NUMBER 99.9)
     32*  MEAN ANNUAL TEMP (DECIMAL NUMBER 99.9)
     33*  MEAN PRECIP JAN (NON-KEY  DECIMAL NUMBER 99.99)
     34*  MEAN PRECIP JUL (NON-KEY  DECIMAL NUMBER 99.99)
     35*  MEAN ANNUAL PREC (DECIMAL NUMBER 99.99)
     36*  HOURS SUN JAN (NON-KEY  INTEGER NUMBER 999)
     37*  HOURS SUN JUL (NON-KEY  INTEGER NUMBER 999)
     38*  REL HUMIDITY JAN (NON-KEY  DECIMAL NUMBER .99)
     39*  REL HUMIDITY JUL (NON-KEY  DECIMAL NUMBER .99)
     40*  CNTY ELEVATION (INTEGER NUMBER 9(5))
     41*  CNTY LATITUDE (NON-KEY  INTEGER NUMBER 9999)
     42*  % WMALE IN MIG (DECIMAL NUMBER 99.9)
     43*  % WMALE OUT MIG (DECIMAL NUMBER 99.9)
     44*  % WFEM IN MIG (DECIMAL NUMBER 99.9)
     45*  % WFEM OUT MIG (DECIMAL NUMBER 99.9)
     46*  % NWMALE IN MIG (DECIMAL NUMBER 99.9)
     47*  % NWMALE OUT MIG (DECIMAL NUMBER 99.9)
     48*  % NWFEM IN MIG (DECIMAL NUMBER 99.9)
     49*  % NWFEM OUT MIG (DECIMAL NUMBER 99.9)
    100*  NEDS EMISSIONS (RG)
      101*  NEDS POLLUTANT (INTEGER NUMBER 9(5) IN 100)
      102*  COUNTY TOTAL (NON-KEY  DECIMAL NUMBER 9(8).999 IN 100)
      103*  POINT SOURCES (NON-KEY  DECIMAL NUMBER 9(8).999 IN 100)
```

(continued)

TABLE 1 (contined)

```
200*  SAROAD MONITORING SITES (RG)
  201*  AREA CODE (INTEGER NUMBER 9999 IN 200)
  202*  SITE CODE (NON-KEY  INTEGER NUMBER 999 IN 200)
  203*  AGENCY (NAME X IN 200)
  204*  PURPOSE (INTEGER NUMBER 99 IN 200)
  205*  CITY (NON-KEY  NAME X(23) IN 200)
  206*  CITY POPULATION (NON-KEY  INTEGER NUMBER 9(8) IN 200)
  207*  UTM ZONE (NAME XX IN 200)
  208*  EASTING (NAME X(8) IN 200)
  209*  NORTHING (NAME X(7) IN 200)
  210*  ADDRESS (NON-KEY  NAME X(41) IN 200)
  211*  TYPE (NAME XX IN 200)
  212*  ELEV ABOVE GND (INTEGER NUMBER 999 IN 200)
  213*  ELEV ABOVE SEA (INTEGER NUMBER 9999 IN 200)
  300*  MONITORING DATA (RG IN 200)
    301*  POLLUTANT (INTEGER NUMBER 9(5) IN 300)
    302*  METHOD (NAME XX IN 300)
    303*  INTERVAL (NAME X IN 300)
    304*  # OBSERVATIONS (NON-KEY  INTEGER NUMBER 9(5) IN 300)
    305*  GEOMETRIC MEAN (DECIMAL NUMBER 9(5).99 IN 300)
    306*  STD DEVIATION (NON-KEY  DECIMAL NUMBER 999.99 IN 300)
    307*  70TH PERCENTILE (NON-KEY  DECIMAL NUMBER 9(5).99 IN 300)
    308*  90TH PERCENTILE (NON-KEY  DECIMAL NUMBER 9(5).99 IN 300)
    309*  99TH PERCENTILE (NON-KEY  DECIMAL NUMBER 9(5).99 IN 300)
    310*  HIGHEST VALUE (DECIMAL NUMBER 9(5).99 IN 300)
    311*  2ND HIGHEST VAL (DECIMAL NUMBER 9(5).99 IN 300)
    312*  LOWEST VALUE (NON-KEY  DECIMAL NUMBER 9(5).99 IN 300)
400*  AIR QUALITY (RG)
  401*  AQ POLLUTANT (INTEGER NUMBER 9(5) IN 400)
  402*  MEAN POP SITES (DECIMAL NUMBER 9(5).99 IN 400)
  403*  MEAN SRCE SITES (DECIMAL NUMBER 9(5).99 IN 400)
  404*  MEAN BKGND SITES (DECIMAL NUMBER 9(5).99 IN 400)
  405*  MEAN TSP PREDICT (DECIMAL NUMBER 9(5).99 IN 400)
600*  PERCENT IN-OUT MIGRATION (RG)
  601*  AGE GROUP (INTEGER NUMBER 99 IN 600)
  602*  % MALE IN TOT (DECIMAL NUMBER 99.99 IN 600)
  603*  % MALE IN ALC (NON-KEY  DECIMAL NUMBER 99.99 IN 600)
  604*  % MALE OUT TOT (DECIMAL NUMBER 99.99 IN 600)
  605*  % MALE OUT ALC (NON-KEY  DECIMAL NUMBER 99.99 IN 600)
  606*  % FEM IN TOT (DECIMAL NUMBER 99.99 IN 600)
  607*  % FEM IN ALC (NON-KEY  DECIMAL NUMBER 99.99 IN 600)
  608*  % FEM OUT TOT (DECIMAL NUMBER 99.99 IN 600)
  609*  % FEM OUT ALC (NON-KEY  DECIMAL NUMBER 99.99 IN 600)
  610*  % WMALE IN TOT (DECIMAL NUMBER 99.99 IN 600)
  611*  % WMALE IN ALC (NON-KEY  DECIMAL NUMBER 99.99 IN 600)
  612*  % WMALE OUT TOT (DECIMAL NUMBER 99.99 IN 600)
  613*  % WMALE OUT ALC (NON-KEY  DECIMAL NUMBER 99.99 IN 600)
  614*  % WFEM IN TOT (DECIMAL NUMBER 99.99 IN 600)
  615*  % WFEM IN ALC (NON-KEY  DECIMAL NUMBER 99.99 IN 600)
  616*  % WFEM OUT TOT (DECIMAL NUMBER 99.99 IN 600)
  617*  % WFEM OUT ALC (NON-KEY  DECIMAL NUMBER 99.99 IN 600)
```

(continued)

TABLE 1 (continued)

```
   618*   % NWMALE IN TOT (DECIMAL NUMBER 99.99 IN 600)
   619*   % NWMALE IN ALC (NON-KEY  DECIMAL NUMBER 99.99 IN 600)
   620*   % NWMALE OUT TOT (DECIMAL NUMBER 99.99 IN 600)
   621*   % NWMALE OUT ALC (NON-KEY  DECIMAL NUMBER 99.99 IN 600)
   622*   % NWFEM IN TOT (DECIMAL NUMBER 99.99 IN 600)
   623*   % NWFEM IN ALC (NON-KEY  DECIMAL NUMBER 99.99 IN 600)
   624*   % NWFEM OUT TOT (DECIMAL NUMBER 99.99 IN 600)
   625*   % NWFEM OUT ALC (NON-KEY  DECIMAL NUMBER 99.99 IN 600)
700*  AGE-ADJUSTED DEATH RATES (RG)
   701*  CAUSE OF DEATH (INTEGER NUMBER 99 IN 700)
   702*  ADJ TOTAL D-RATE (DECIMAL NUMBER 9999.99 IN 700)
   703*  ADJ WMALE D-RATE (DECIMAL NUMBER 9999.99 IN 700)
   704*  ADJ WFEM D-RATE (DECIMAL NUMBER 9999.99 IN 700)
   705*  ADJ NWMAL D-RATE (DECIMAL NUMBER 9999.99 IN 700)
   706*  ADJ NWFEM D-RATE (DECIMAL NUMBER 9999.99 IN 700)
800*  AGE-SPECIFIC POPULATION (RG)
   801*  AGE GROUPING (INTEGER NUMBER 99 IN 800)
   802*  TOTAL POP (NON-KEY  INTEGER NUMBER 9(8) IN 800)
   803*  FEMALE (NON-KEY  INTEGER NUMBER 9(8) IN 800)
   804*  WHITE (NON-KEY  INTEGER NUMBER 9(8) IN 800)
   805*  WHITE FEMALE (NON-KEY  INTEGER NUMBER 9(8) IN 800)
   806*  NONWHITE MALE (NON-KEY  INTEGER NUMBER 9(8) IN 800)
   807*  TOTAL URBAN POP (NON-KEY  INTEGER NUMBER 9(8) IN 800)
   808*  URBAN FEMALE (NON-KEY  INTEGER NUMBER 9(8) IN 800)
   809*  URBAN WHITE (NON-KEY  INTEGER NUMBER 9(8) IN 800)
   810*  URB WHITE FEMALE (NON-KEY  INTEGER NUMBER 9(8) IN 800)
```

User-defined functions:

```
1001*  AREA SOURCES (DECIMAL FUNCTION ((C102-C103)))
1002*  MALE (INTEGER FUNCTION ((C802-C803)))
1003*  NONWHITE (INTEGER FUNCTION ((C802-C804)))
1004*  WHITE MALE (INTEGER FUNCTION ((C804-C805)))
1005*  NONWHITE FEMALE (INTEGER FUNCTION ((C803-C805)))
```

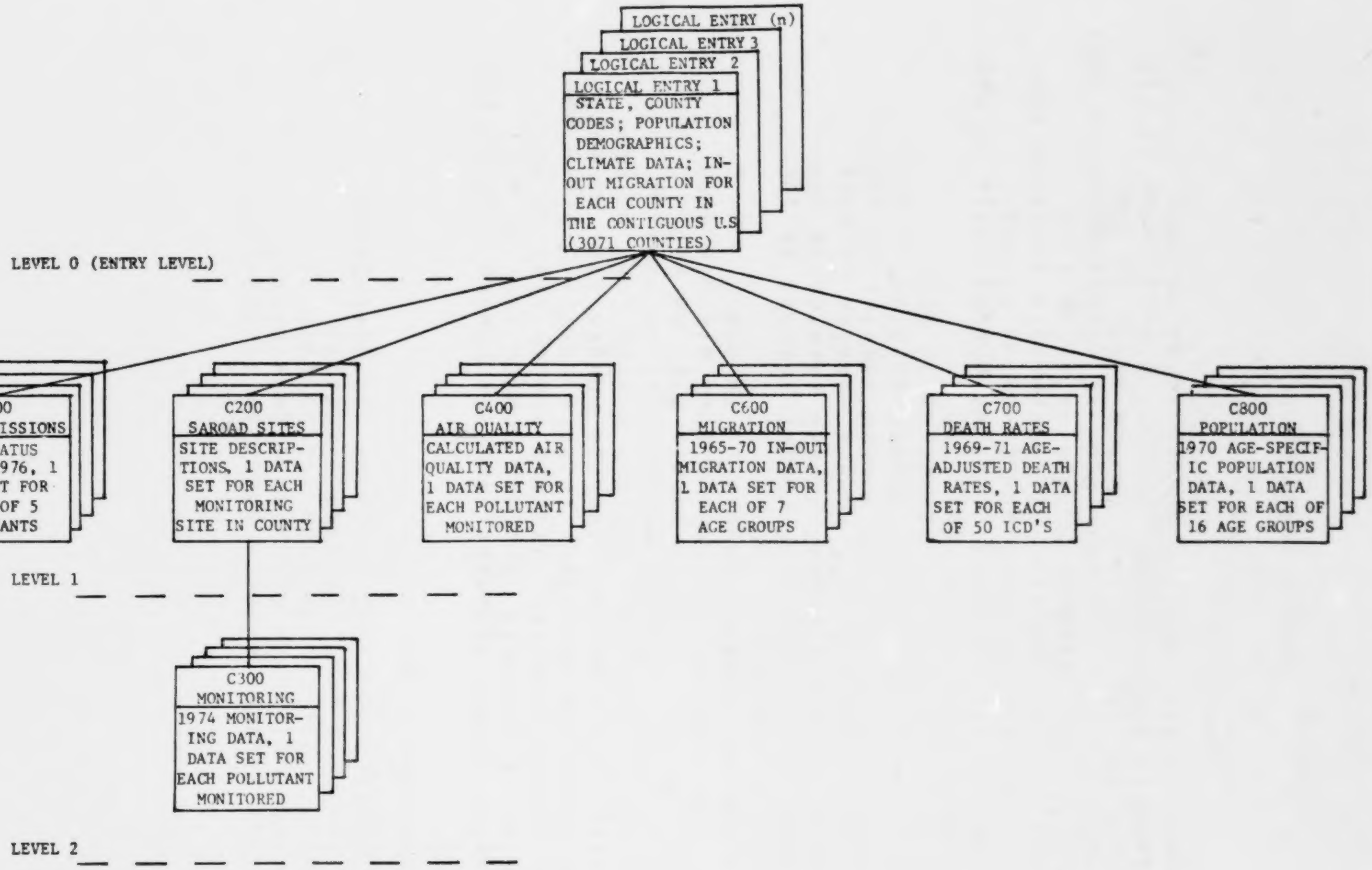

LOGICAL ENTRY (n)
LOGICAL ENTRY 3
LOGICAL ENTRY 2
LOGICAL ENTRY 1
STATE, COUNTY CODES; POPULATION DEMOGRAPHICS; CLIMATE DATA; IN-OUT MIGRATION FOR EACH COUNTY IN THE CONTIGUOUS U.S (3071 COUNTIES)
LEVEL 0 (ENTRY LEVEL)
C100
NEDS EMISSIONS
FILE STATUS AS OF 1976, 1 DATA SET FOR EACH OF 5 POLLUTANTS
C200
SAROAD SITES
SITE DESCRIPTIONS, 1 DATA SET FOR EACH MONITORING SITE IN COUNTY
C400
AIR QUALITY
CALCULATED AIR QUALITY DATA, 1 DATA SET FOR EACH POLLUTANT MONITORED
C600
MIGRATION
1965-70 IN-OUT MIGRATION DATA, 1 DATA SET FOR EACH OF 7 AGE GROUPS
C700
DEATH RATES
1969-71 AGE-ADJUSTED DEATH RATES, 1 DATA SET FOR EACH OF 50 ICD'S
C800
POPULATION
1970 AGE-SPECIFIC POPULATION DATA, 1 DATA SET FOR EACH OF 16 AGE GROUPS
LEVEL 1
C300
MONITORING
1974 MONITORING DATA, 1 DATA SET FOR EACH POLLUTANT MONITORED
LEVEL 2

Figure 1. POPATRISK heirarchical structure.

descendant of C200 SAROAD MONITORING SITES and thus, entry level data 1* through 41*, repeating group C200, and repeating group C300 are related and belong to the same family tree. Logically, this means that for a county C2 there are a number of data sets in C200 containing information on each monitoring site for the county, and for each monitoring site data set, there are a number of data sets in C300 containing information on each pollutant monitored by each site.

COMPONENTS AND SOURCES OF DATA

Table 2 is a quick-reference description of each component included in POPATRISK. Each component number is listed in numerical data base definition order along with the component name, description, units, picture in COBOL notation, and original source of the data item. Certain components in the data base definition (Table 1) are described as 'NON-KEY.' All unlabelled components are 'KEY' except those labelled RG (repeating group). Key components in Table 2 are flagged with a + symbol to the left of the component number. It is anticipated that these KEY items will be frequently needed as selection criteria for other data in the data base. A component must be KEY to be used in a retrieval WHERE clause. For example, for KEY component C305, one could easily obtain the number of stations reporting a yearly geometric mean greater than a specified value for any pollutant, for any set of counties. Because of the significant cost of loading and maintaining KEY'ed components, the KEY option is used only for those components considered useful as selection criteria.

State and county codes (components C1 through C3) conform to the SAROAD county assignment protocol except for isolated cases which are discussed later in this section and also in the 'POPATRISK Geocoding Manual.' General county description data (components C10 through C41) were extracted from a master file created by SSI. Population figures are based on the 1970 Census and, with the exception of TOTAL POPULATION, were extracted from a Bureau of the Census 1972 County and City Data Book computer file containing data obtained from representative samples of population rather than a complete count. To maintain consistency as much as possible and to prevent possible confusion, TOTAL POPULATION C10 figures were changed to reflect the same complete

TABLE 2. POPATRISK DATA BASE COMPONENT DESCRIPTION

COMPONENT NUMBER	COMPONENT NAME*	COMPONENT DESCRIPTION	DATA VALUE TYPE/LENGTH**	SOURCE OF DATA***
+ 1	STATE ID	POPATRISK state code (SAROAD code for contiguous United States) as in SAROAD station coding manual. EPA Office of Air Programs Publication # APTD-0907.	XX	SAROAD
+ 2	COUNTY ID	POPATRISK county code (modified SAROAD coding scheme as discussed in POPATRISK Geocoding Manual).	XXXX	SAROAD - SSI
+ 3	ST-COUNTY ID	State-county combined code, a unique code for each county (logical entry).	X(6)	SAROAD-SSI
+ 10	TOTAL POPULATION	Total county population as determined from 1970 Census. Data are available, e.g., in hard copy, in Bureau of the Census Publication PC(V1), Final Population Counts.	9(8)	Census
+ 11	PERCENT NONWHITE	Percent of total population that are non-white as defined by Census.	99.9	SSI-County and City Data Book, 1972
+ 12	PERCENT BLACK	Percent black in total population	99.9	"
+ 13	PERCENT OVER 64	Percent of total population over 64 years old	99.9	"
+ 14	PERCENT FEMALE	Percent female of total population	99.9	"
+ 15	POP PER SQ MI	Population per square mile in county	9(5).9	"
+ 16	PERCENT URBAN	Percent of total county population that is urban as defined by Census.	99.9	"
17	% EMP OTHER CNTY	Percent of population employed in another county	99.9	"
18	INCOME >10K/<5K	Ratio of families with income greater than $10,000 to families with income less than $5,000, 1969.	99.99	"
19	INCOME >15K/<10K	Ratio of families with income greater than $15,000 to families with income less than $10,000, 1969.	99.99	"
20	INCOME >15K/<5K	Ratio of families with income greater than $15,000 to families with income less than $5,000, 1969.	99.99	"
21	% USE PUB TRANS	Percent of population that uses public transportation to get to work	99.9	"
+ 22	ALCOHOL SALE >18	Total sales for liquor stores divided by population of county over 18 years of age. Calculated using data from 1972 Census of Business, Retail Trade (Dollars).	999.9	SSI-Census of Business, Retail Trade
+ 23	GAS SALES/CAPITA	Gas Sales per capita, measured by the value of total sales by gasoline stations (Dollars).	999.9	"

+ Denotes component is KEY.

* Component names are 16 characters or less to provide full default headings on LIST retrievals.

** Standard COBOL field description format.

*** See Section 2 for further documentation of sources.

(continued)

TABLE 2 (continued)

COMPONENT NUMBER	COMPONENT NAME*	COMPONENT DESCRIPTION	DATA VALUE TYPE/LENGTH**	SOURCE OF DATA***
30	MEAN TEMP JAN	Mean temperature for month of January (°F)	99.9	SSI-NOAA
31	MEAN TEMP JUL	Mean temperature for month of July (°F)	99.9	"
+ 32	MEAN ANNUAL TEMP	Mean annual temperature (°F)	99.9	"
33	MEAN PRECIP JAN	Mean precipitation for month of January (Inches)	99.99	"
34	MEAN PRECIP JUL	Mean precipitation for month of July (Inches)	99.99	"
+ 35	MEAN ANNUAL PREC	Mean annual precipitation (Inches)	99.99	"
36	HOURS SUN JAN	Mean hours of sun for month of January (Hours)	999	"
37	HOURS SUN JUL	Mean hours of sun for month of July (Hours)	999	"
38	REL HUMIDITY JAN	Mean relative humidity for month of January	.99	"
39	REL HUMIDITY JUL	Mean relative humidity for month of July	.99	"
+ 40	CNTY ELEVATION	County elevation (Feet)	9(5)	"
41	CNTY LATITUDE	County latitude (Degrees-Minutes)	9999	"
+ 42	% WMALE IN MIG	Percent of county white male population that are in-migrants	99.9	SSI-Census Special Tape
		Migration data were calculated from a specially tabulated 1965-70 migration file provided by the Bureau of the Census. Migration totals are expressed as the percent of a particular race/sex group that were migrants, e.g., % WMALE IN MIG is total white male in-migrants for the county for the years 1966, 1967, 1968, 1969, 1970/total white males in the county as of 1970.		
+ 43	% WMALE OUT MIG	Percent white male out-migrants	99.9	"
+ 44	% WFEM IN MIG	Percent white female in-migrants	99.9	"
+ 45	% WFEM OUT MIG	Percent white female out-migrants	99.9	"
+ 46	% NWMALE IN MIG	Percent non-white male in-migrants	99.9	"
+ 47	% NWMALE OUT MIG	Percent non-white male out-migrants	99.9	"
+ 48	% NWFEM IN MIG	Percent non-white female in-migrants	99.9	"
+ 49	% NWFEM OUT MIG	Percent non-white female out-migrants	99.9	"
100	NEDS EMISSIONS	NEDS emissions data repeating group containing emissions totals for 5 criteria pollutants. Data were extracted from the NADB-NEDS-USER file and are current as of the 1976 data file status.		
+101	NEDS POLLUTANT	Five digit code for NEDS pollutant--11101 (TSP), 42101 (Carbon Monoxide), 42401 (Sulfur Dioxide), 42602 (Nitrogen Dioxide), 43101 (Hydrocarbons)	9(5)	NEDS-USER file

(continued)

TABLE 2 (continued)

COMPONENT NUMBER	COMPONENT NAME*	COMPONENT DESCRIPTION	DATA VALUE TYPE/LENGTH**	SOURCE OF DATA***
102	COUNTY TOTAL	Total emissions for the county by pollutant. (Tons/yr)	9(8).999	
103	POINT SOURCES	Point source emissions for the county by pollutant. (Tons/yr)	9(8).999	
		SAROAD site descriptions and monitoring data for each pollutant monitored at each site. Extracted from the 1974 NADB SAROAD files: Yearly Summary File (NADB-YRSUMS-D), Frequency File (NADB -YRFRQ-D), and Site Description File (NADB-STE-INV).		
200	SAROAD MONITORING SITES	1974 SAROAD site descriptions for each air quality monitoring site in the county. Codes may be found in the Directory of Air Quality Monitoring Sites Active in 1974. EPA-450/2-76-008.		NADB-STE-INV
+ 201	AREA CODE	City, county, parish or air pollution control district	9999	"
202	SITE CODE	Unique sampling site ID within city, county, parish or air pollution district	999	"
+ 203	AGENCY	Agency type, code A-Z as shown in Table 1, Directory of Air Quality Monitoring Sites active in 1974. EPA-450/2-76-008.	X	"
+ 204	PURPOSE	Project classification, Code 01-99 as shown in Table 2 of above publication	99	"
205	CITY	City where monitor is located	X(23)	"
206	CITY POPULATION	Population of city monitored	9(8)	"
+ 207	UTM ZONE	One of 10 Universal Transverse Mercator zones covering the continental U.S.	XX	"
+ 208	EASTING	Horizontal coordinate of monitor within UTM zone (Meters)	X(8)	"
+ 209	NORTHING	Distance of monitor from the equator (Meters)	X(7)	"
210	ADDRESS	Site address	X(41)	"
+ 211	TYPE	Station type, as described in Vol. 5, AEROS Manual of Codes EPA-450/2-76-005	XX	"
+ 212	ELEV ABOVE GND	Elevation of monitor above ground. (Feet)	999	"
+ 213	ELEV ABOVE SEA	Elevation of monitor above mean sea level. (Feet)	9999	"
300	MONITORING DATA	1974 SAROAD monitoring data for each pollutant monitored at the site		

(continued)

TABLE 2 (continued)

COMPONENT NUMBER	COMPONENT NAME*	COMPONENT DESCRIPTION	DATA VALUE TYPE/LENGTH**	SOURCE*** OF DATA
+ 301	POLLUTANT	Five digit code for SAROAD pollutant monitored 11101 (TSP), 12306 (Nitrate), 12403 (Sulfate), 42101 (Carbon Monoxide), 42269 (Total Sulfur) 42401 (Sulfur dioxide), 42602 (Nitrogen dioxide), 44201 (Ozone)	9(5)	NADB-YRFREQ-D
+ 302	METHOD	Data collection method code as defined in Vol. 5, AEROS Manual of Codes. EPA-450/2-76-005	XX	"
+ 303	INTERVAL	Sampling interval, as described in above publication	X	"
304	# OBSERVATIONS	Number of observations of pollutant concentration made in 1974 for given sampling interval.	9(5)	"
+ 305	GEOMETRIC MEAN	Geometric mean of measurements in 1974. ($\mu g/m^3$)	9(5).99	NADB-YRSUM-D
306	STD DEVIATION	Geometric standard deviation from the 1974 mean. ($\mu g/m^3$)	999.99	"
307	70TH PERCENTILE	Observed measurement below which 70% of the values fall. (μ/m^3)	9(5).99	NADB-YRFREQ-D
308	90TH PERCENTILE	Observed measurement below which 90% of the values fall. ($\mu g/m^3$)	9(5).99	"
309	99TH PERCENTILE	Observed measurement below which 99% of the values fall. ($\mu g/m^3$)	9(5).99	"
+ 310	HIGHEST VALUE	Highest value observed during 1974. ($\mu g/m^3$)	9(5).99	"
+ 311	2ND HIGHEST VAL	2nd highest value observed during 1974. ($\mu g/m^3$)	9(5).99	"
312	LOWEST VALUE	Lowest value observed during 1974. ($\mu g/m^3$)	9(5).99	NADB-YRSUM-D
400	AIR QUALITY	Repeating group containing arithmetic mean of values of monitoring data for each SAROAD pollutant monitored for purpose codes of 01, 02, and 03, where sampling interval equals 1 or 7.		
+ 401	AQ POLLUTANT	Five digit code for SAROAD pollutant monitored (code same as for component 301)	9(5)	"
+ 402	MEAN POP SITES	Arithmetic mean of the geometric means for each county site for each pollutant where purpose code equals 01 (Population-oriented surveillance) and where the sampling interval equal 1 (1 hour) or 7 (24 hours). ($\mu g/m^3$)	9(5).99	SSI

(continued)

TABLE 2 (continued)

COMPONENT NUMBER	COMPONENT NAME*	COMPONENT DESCRIPTION	DATA VALUE TYPE/LENGTH**	SOURCE OF DATA***
+ 403	MEAN SOURCE SITES	Arithmetic mean of the geometric means for each county site for each pollutant where purpose code equals 02 (Source-oriented ambient surveillance), and where the sampling interval equals 1 or 7. ($\mu g/m^3$)	9(5).99	SSI
+ 404	MEAN BKGND SITES	Arithmetic mean of the geometric means for each county site for each pollutant where purpose code equals 03 (Background surveillance), and where the sampling interval equals 1 or 7. ($\mu g/m^3$)	9(5).99	SSI
+ 405	MEAN TSP PREDICT	Mean predicted TSP monitoring, calculated for counties where TSP and NO_x emissions fall within a specific range. Emission ranges and the algorithm for calculating the mean predicted TSP are discussed in Section 2. ($\mu g/m^3$)	9(5).99	SSI
600	PERCENT IN-OUT MIGRATION	County in-out migration repeating group containing data based on 1970 residence with regard to residence in 1965, expressed as a percentage of the total county population for each race-sex group. Data is provided in terms of Total (Allocated + Unallocated) and Allocated. An unallocated group is defined by actual answers to a 1970 Census question. An allocated group is defined by a Census algorithm. (See also component 42.) Census migration tape documentation (Appendix B) provides a discussion of the migration data and allocation procedure.		SSI-Census Special Tape
+ 601	AGE GROUP	A numerical code defining the age span included in the group and consisting of the last age in the age span for the group, e.g. 14(5-14), 19(15-19), 24(20-24), 29(25-29), 44(30-44), 64(45-64), 65(65 and older).	99	"
+ 602	% MALE IN TOT	Percent of the total males in the county that migrated in and are in the corresponding age group. Total = Allocated + Unallocated.	99.99	"
603	% MALE IN ALC	Percent male in (allocated).	"	"
+ 604	% MALE OUT TOT	Percent male out (total).	"	"
605	% MALE OUT ALC	Percent male out (allocated).	"	"
+ 606	% FEM IN TOT	Percent female in (total).	"	"
607	% FEM IN ALC	Percent female in (allocated).	"	"
+ 608	% FEM OUT TOT	Percent female out (total).	"	"
609	% FEM OUT ALC	Percent female out (allocated).	"	"

(continued)

TABLE 2 (continued)

COMPONENT NUMBER	COMPONENT NAME*	COMPONENT DESCRIPTION	DATA VALUE TYPE/LENGTH**	SOURCE OF DATA***
+ 610	% WMALE IN TOT	Percent white male in (total).	99.99	Census
611	% WMALE IN ALC	Percent white male in (allocated).	"	"
+ 612	% WMALE OUT TOT	Percent white male out (total).	"	"
613	% WMALE OUT ALC	Percent white male out (allocated).	"	"
+ 614	% WFEM IN TOT	Percent white female in (total).	"	"
615	% WFEM IN ALC	Percent white female in (allocated).	"	"
+ 616	% WFEM OUT TOT	Percent white female out (total).	"	"
617	% WFEM OUT ALC	Percent white female out (allocated).	"	"
+ 618	% NWMALE IN TOT	Percent non-white male in (total).	"	"
619	% NWMALE IN ALC	Percent non-white male in (allocated).	"	"
+ 620	% NWMALE OUT TOT	Percent non-white male out (total).	"	"
621	% NWMALE OUT ALC	Percent non-white male out (allocated).	"	"
+ 622	% NWFEM IN TOT	Percent non-white female in (total).	"	"
623	% NWFEM IN ALC	Percent non-white female in (allocated).	"	"
+ 624	% NWFEM OUT TOT	Percent non-white female out (total).	"	"
625	% NWFEM OUT ALC	Percent non-white female out (allocated).	"	"
700	AGE-ADJUSTED DEATH RATES	Age-adjusted mortality repeating group containing 1969-71 combined age-adjusted death rates for the 48 ICDA aggregated categories + 2 malignant neoplasm cluster categories. (See Table 3 and Table 4.)		
+ 701	CAUSE OF DEATH	Code 1-50 for each aggregation of ICDA categories	99	NCHS
+ 702	ADJ TOTAL D-RATE	Total age-adjusted death rate for each ICDA category (per 100,000).	9999.99	NCHS
+ 703	ADJ WMALE D-RATE	Age-adjusted white male death rate for each aggregation of ICDA categories (per 100,000).	"	"
+ 704	ADJ WFEM D-RATE	Age-adjusted white female death rate (per 100,000).	"	"
+ 705	ADJ NWMALE D-RATE	Age-adjusted non-white male death rate (per 100,000).	"	"
+ 706	ADJ NWFEM D-RATE	Age-adjusted non-white female death rate (per 100,000).	"	"

(continued)

TABLE 2 (continued)

COMPONENT NUMBER	COMPONENT NAME*	COMPONENT DESCRIPTION	DATA VALUE TYPE/LENGTH**	SOURCE OF DATA***
800	AGE-SPECIFIC POPULATION	Age-specific population data repeating group containing 1970 Census data grouped into 15 age categories. A 16th category is provided with population totals across the age groups. Demographic groupings included provide figures for total population and urban population. An appropriate definition of urban may be found in the 1972 County and City Data Book.		
+ 801	AGE GROUPING	A numerical code defining the age span included in the group and consisting of the last age in the age span for the group. There are 16 age groups, 14 of which cover 5 year increments in age from 0 to 69 such that age group 4 covers ages 0-4, 9 covers 5-9, and so on. A 15th age group group, grouping 99, covers ages 70 and over. The 16th age group, grouping 0, represents population totals for all age groups.	99	Census
802	TOTAL POP	Total population for each age group	9(8)	Census
803	FEMALE	Female population for each age group	"	"
804	WHITE	White population for each age group	"	"
805	WHITE FEMALE	White female population for each age group	"	"
806	NOWHITE MALE	Non-white male population for each age group	"	"
807	TOTAL URBAN POP	Total urban population for each age group	"	"
808	URBAN FEMALE	Urban female population for each age group	"	"
809	URBAN WHITE	Urban white population for each age group	"	"
810	URB WHITE FEMALE	Urban white female population for each age group		
USER DEFINED FUNCTIONS:				
C1001	AREA SOURCES	Area source emissions for county (Tons/yr) (C102-C103)		
C1002	MALE	Male population for age group specified (C802-C803)		
C1003	NONWHITE	Non-white population for age group specified (C802-C804)		
C1004	WHITE MALE	White male population for age group specified (C804-C805)		
C1005	NONWHITE FEMALE	Non-white female population for age group specified (C803-C805)		

count as is used in the age-specific population repeating group. Income ratios are based on 1969 data and were also taken from the 1972 County and City Data Book file.

Figures for alcohol sales and gasoline sales were taken from the 1970 Census of Retail Trade. Data for alcohol sales are available only when 500 or more retail establishments exist in the county. Therefore, data is available for roughly 60 to 70 percent of the counties. The problem is further complicated by disclosure limitations in some of the remaining counties. Using the data that do exist, SSI developed an index for providing a rough indication of alcohol sales per persons 18 years of age and over for unlisted counties.

Climate data were originally obtained from the Environmental Data Service of the National Oceanic and Atmospheric Administration. The data are in terms of means based on the years 1941-70. Where there is more than one weather station, data are recorded for the station closest to the county's population center. The same procedure was used to record the county's elevation and latitude. Since not all counties have a primary weather station, National Weather Service climate regions were used to extrapolate data from other similar climatic areas.

County migration totals (components C42 through C49) are totals for in and out migration summed across the age groups in the migration repeating group. Detailed information on the in-out migration data will be provided later in this section.

The NEDS EMISSIONS repeating group data (components C101 through C103) were extracted from data present in the NADB NEDS-USER file in 1976 for the five criteria pollutants: Total Suspended Particulates(11101), Carbon Monoxide(42101), Sulfur Dioxide(42401), Nitrogen Dioxide(42602), and Hydrocarbon(43101). The numbers in parentheses refer to standard NADB pollutant codes. The repeating group consists of data sets containing total county emissions and total point source emissions for each of the pollutants. A user-defined function is provided for calculation of area source emissions by subtracting the point source emissions from total emissions.

The SAROAD MONITORING SITES data (components C201 through C312) were selected from 1974 data in three SAROAD files; Yearly Summary File (NADB-YRSUM-D), Frequency File (NADB-YRFRQ-D), and the Site Description File (NADB-STE-INVO). The SAROAD MONITORING SITES repeating group contains data sets describing each monitoring site, its location, type, and purpose. The MONITORING DATA repeating group consists of data sets containing SAROAD monitoring data as descendants of the reporting site descriptions.

Monitoring data are provided for the following pollutants: TSP(11101), Nitrate(12306), Sulfate(12403), Carbon Monoxide(42101), Total Sulfur(42269), Sulfur Dioxide(42401), Nitrogen Dioxide(42602), and Ozone(44201). The numbers in parentheses refer to standard NADB pollutant codes. Sampling intervals provided in the data base are identical to standard NADB codes, except that codes X, Y, Z, and C were expunged from the data base because the data were of questionable validity.

As requested by the EPA Project Officer, the yearly means in the data base are geometric means. Although some present National Air Quality Standards set a limit on annual arithmetic means, these means may be estimated from a knowledge of the geometric means and geometric standard deviations. A list and description of codes used in the data base to describe monitoring sites and data, e.g. site type, sampling method, sampling interval may be found in the appropriate sources as referenced by Table 2.

The AIR QUALITY DATA repeating group data (components C401 through C405) represent calculations of air quality data for each pollutant monitored in the county. Data are in terms of the arithmetic mean of geometric means for each monitored pollutant for purposes: (1) population monitoring, (2) source monitoring, and (3) background monitoring where the sampling interval code is 1 (1 hour) or 7 (24 hour). An additional component provides a calculated predicted mean for TSP where TSP and NO_x emissions (tons/year) fall within the following limits:[1]

$$1414.375 \leq TSP \leq 230031.803$$
$$NO_x \leq 432778.195$$

---

[1] These limits correspond to the limits of the variable set used in the regression analysis discussed in Appendix C.

The PERCENT IN-OUT MIGRATION repeating group data (components C601 through C625) represent calculated migration rates for the county over the span of years 1965-70. Documentation on the in-out migration file, as received from the Bureau of the Census, is provided in Appendix B. It provides a discussion of the file content, allocation procedures, and comparison with published data. Briefly, the data in the file are based on question 19 of the 1970 Census 15 percent sample household questionnaire. It requested those persons who reported living in a different house in 1965 to report the state, county, and city for their residence in 1965. An allocation procedure was developed by the Bureau of the Census to provide a residence assignment procedure for those persons indicating they lived in a different house but did not report an address. Data in the file are in terms of both 'unallocated' and 'allocated' figures. It was decided that for use in the data base, the data would be best represented in terms of Total (allocated + unallocated) and Allocated. The race categories on the data file are white, black, and total all races. To maintain consistency throughout the data base, race categories in the data are white, non-white, and total all races. The 7 age groups used in the migration file roughly correspond to aggregations of the 15 age groups for age-specific population data to be discussed later. These seven age groups are used in the data base.

The in-out migration repeating group in the data base consists of seven data sets, one data set for each of the age groups in the migration file. For each data set there are figures for each of the race-sex groups in terms of Total and Allocated. Since the Total figures are considered to be the most useful, those components are KEY'ed.

The AGE-ADJUSTED DEATH RATES repeating group (components C701 through C706) contains direct method age-adjusted mortality rates for the county over a three year period from 1969 to 1971. Three tapes of raw mortality data, one tape for each year, were obtained from the National Center For Health Statistics. Each tape contains about two million records, one record per death, and includes all deaths for the year. The cause of death contained in each record is coded using a four-digit code of the Eighth Revision of the International Classification of Diseases. Considerations of constraints imposed by data base size limitations and obtaining significant numbers of deaths over a three-year period required some aggregation of mortality categories at the three-digit level. With the approval of the EPA Project Officer, a decision was made to use aggregations of mortality categories corresponding to those already in use at HERL.

Two supplemental malignant neoplasm cluster groups were gleaned from the publication 'A Sequential Space-Time Cluster Analysis of Cancer Mortality in the United States: Etiologic Implications,' Fred Burbank, Am. J. of Epid 95, 393 (1972). Further discussion of the selection methodology for these clusters is provided in Section 5.

Table 3 describes the 50 aggregations of mortality categories used in the POPATRISK data base. Table 4 gives a detailed description of the two cluster groups. The POPATRISK code associated with each cause grouping in Table 3 is the code by which the grouping is referenced in the data base. The repeating group consists of one data set per cause grouping, containing county age-adjusted rates for the total population and four race-sex groups.

The AGE-SPECIFIC POPULATION repeating group (components C801 through C810) contains age-specific population data extracted from 1970 Census tapes. Population counts on the Census tapes are based on a 100% count and were provided for rural and urban and broken down by race-sex group and age. For loading into the data base, the data were aggregated into 15 age categories, 1 data set per category, and expressed in terms of Total and Urban race-sex groups. A 16th data set is provided as totals across the 15 age groups. Race-sex group data not represented in the repeating group are provided for as user-defined functions.

The '1000' series components are user-defined functions which may be changed or augmented at any time without affecting the integrity of the data base. They are mathematical expressions intended to provide useful information needed on a less frequent basis while conserving storage space in the data base.

TABLE 3. AGGREGATIONS OF MORTALITY CATEGORIES FOR POPATRISK AGE-ADJUSTED DEATHS

POPATRISK Cause of Death	POPATRISK Code	NCHS-60 Causes	ICD-7	ICD-8
Tuberculosis	1	00, 01	001-019	010-019
Infective and parasitic diseases	2	02-10	020-138	000-009 020-136
Malignant neoplasm of buccal cavity and pharynx	3	11	140-148	140-149
Malignant neoplasm of digestive organs and peritoneum	4	12	150-156A 157-159	150-159
Malignant neoplasm of respiratory system	5	13	160-164	160-163
Malignant neoplasm of breast	6	14	170	174
Malignant neoplasm of genital organs	7	15	171-179	180-187
Malignant neoplasm of urinary organs	8	16	180-181	188-189
Malignant neoplasm of other unspecified sites	9	17	156B,165 190-199	170-173 190-199
Leukemia and aleukemia	10	18	204	204-207
Lymphosarcoma and other neoplasms of lymphatic and hematoporetic tissue	11	19	200-203 205	200-203 208-209
Benign neoplasms of unspecified nature	12	20	210-238	210-239
Asthma	13	21	241	493
Diabetes mellitus	14	22	260	250
Anemias	15	23	290-293	280-285
Meningitis	16	24	340	320
Cerebrovascular disease	17	25	330-334	430-438
Rheumatic fever and chronic rheumatic heart disease	18	26	400-402 410-416	390-398
Arteriosclerotic heart disease	19	27	420	410-414
Chronic diseased of endocordium	20	28	421,422	424-429
Other forms of heart disease	21	29	430-434	420-423
Hypertensive heart disease	22	30	440-443	402-404
Other hypertensive disease	23	31	444-447	400-401
General arteriosclerosis	24	32	450	440
Other diseases of circulatory system	25	33	451-468	441-458
Chronic and unspecified nephritis and other renal sclerosis	26	34	592-594	582-584

(continued)

TABLE 3 (continued)

POPATRISK Cause of Death	POPATRISK Code	NCHS-60 Causes	ICD-7	ICD-8
Influenza and pneumonia	27	35,36	480-493	470-486
Bronchitis	28	37	500-502	490-492
Other bronchopulmonic disease	29	38	510-527 470-475	500-519 460-466
Ulcer of stomach and duodenum	30	39	540-541	531-532
Appenidicts, hernia and intestinal obstruction	31	40,41	550-570	540-543 550-553 560
Gastrites, duodenitis, enteritis, and colitic, except diarrhea of newborn	32	42	543,571, 572	535 561-563
Cirrhosis of liver	33	43	581	571
Cholelithiasis, cholecystitis, and cholangitis	34	44	584-585	574-575
Nephritis and kidney infection	35	45,46	590,591, 600	580-581 590
Hyperplasia of prostate	36	47	610	600
Deliveries and complication of pregnancy	37	48,49	640-689	630-678
Congenital anomalies	38	50	750-759	740-759
Birth injuries and diseases	39	51,52	760-768	760-773
Other diseases of early infancy	40	53	769-776	774-779
Symptoms, senility, and ill-defined conditions	41	54	780-795	780-796
All other diseases	42	55	Residual	Residual
Motor vehicle accidents	43	56	E810-E835	E810-E823
Other accidents	44	57	E800-E802 E840-E962	E800-E807 E825-E949
Suicide	45	58	E963,E970-E979	E950-E959
Homicide	46	59	E964,E980-E985	E960-E969
Other injuries	47			E970-E999
All causes	48	1-59	1-999	1-999
Malignant neoplasm cluster 1	49			151-154, 17 187-188
Malignant neoplasm cluster 2	50			141, 147 161-163

TABLE 4. CLUSTER DESCRIPTIONS

Cluster 1:	ICD	Rates Male *	Rates Female *
Breast	174	0.24	20.89
Bladder	188	4.89	1.90
Kidney	187	3.25	1.81
Rectum	154	4.71	3.44
Stomach	151	12.93	6.64
Large Intestine	153	12.29	13.67
Small Intestine	152	?	?
Cluster 2:			
Tongue	141	?	?
Naso Pharynx	147	0.33	0.10
Trachea, Bronchus	162	29.59	5.07
and Lung	163		
Larynx	161	2.0	0.20

* Mortality rates cited are 50th percentile values taken from Atlas of Cancer Mortality for U.S. Counties: 1950-1969, U.S. Department of Health, Education and Welfare. DHEW Publication No. (NIH)75-780.

VARIATIONS IN GEOCODING SCHEMES AMONG SOURCES

The compilation of several different types of data originating from a variety of independent sources to create POPATRISK resulted in several discrepancies in county code assignment schemes. Resolution of these discrepancies involved a significant portion of the data processing effort. A brief description of the geocoding problems and their impact on the POPATRISK coding scheme will be outlined here while a detailed discussion of the resolution methodology is provided in Section 5. Full documentation of the POPATRISK geocoding scheme, including a cross-reference table to the various other county codes is provided under separate cover in the POPATRISK Geocoding Manual.

The state of Massachusetts presented a special problem when using SAROAD county assignments in the POPATRISK data base because the county assignments do not match those of any of the other data sources. SAROAD recognizes 13 counties while NEDS has 6 counties and Census recognizes 14 counties. With few exceptions, Census county assignments adhere to the Federal Information Processing Standards (FIPS) coding scheme. Since Census uses a finer geographical breakdown than NEDS and SAROAD and is most widely used, it was decided to adopt the Census protocol for Massachusetts. All data not conforming to Census protocol, namely NEDS and SAROAD, required development of a reallocation procedure to reassign data to the proper counties; Massachusetts county codes to be used in the data base were arbitrarily chosen to be the Census county codes with a leading zero to fulfill the four-digit code requirement.

Throughout development of the data base, Virginia continued to be a source of problems when attempting to mold data for Virginia counties to conform to a county based coding scheme. The independent cities were not compatible with the county based data base and data for each city had to be reassigned to its original county and incorporated into the data for that county entry. It was necessary to "create" two new counties in Virginia for those independent cities for which no county now exists, namely those areas formerly occupied by Warwick County and Elizabeth City, and also by Norfolk and Princess Anne counties. These two new counties were arbitrarily assigned county codes of 0972 and 2144, respectively, to maintain consistency in the numerical and alphabetical order of counties.

Another problem with independent cities occurred for the cities of Baltimore, Maryland and St. Louis, Missouri. Both FIPS and SAROAD lists counties alphabetically but while SAROAD lists Baltimore before Baltimore County, FIPS lists them in opposite order. A similar problem existed with St. Louis, Missouri. Yellowstone, Montana also appeared as an independent city and the data were reassigned to Park County, Montana.

The five New York City boroughs presented a problem in that some data sources provided data only for New York City as a whole with no further breakdown into boroughs as needed for POPATRISK. Climate and geographical data were two such examples. In these cases, the data provided for New York City were duplicated for each of the five boroughs. NCHS mortality data were also available only for New York City as a whole. In this case, a decision was made to calculate age-adjusted death rates using total New York City data along with the total population figures for all five boroughs, yielding rates for New York City as a whole. These rates were then duplicated for each of the five New York City boroughs.

SECTION 3

USE OF DATA BASE "POPATRISK"

The POPATRISK data base resides on a private removable disk pack on the Univac 1110 at Research Triangle Park. It is accessable in either demand or batch processing mode and can be used by non-programmers with a basic knowledge of SYSTEM 2000 Natural Language. As part of SYSTEM 2000's security system, a user is required to give a password for access to the data base. Multiple passwords may be available to the user, but these passwords should have only retrieval authorities. Assignment of passwords and their authorities are allowed only by the Data Base Administrator using his/her master password. Normally the Data Base Administrator maintains exclusive update authority to the data base.

Once in possession of a valid password, a user may access the data base using the following commands:

```
@XQT  N*NS2K.280
USER, 'Password':
DATA BASE NAME IS POPATRISK:
   S2K Natural Language Commands
    '
    '
    '
EXIT:
```

As previously mentioned, it is not within the scope of this report to provide instruction in the use of SYSTEM 2000 Natural Language. However, there are some points to discuss concerning various retrieval and maintenance techniques tailored specifically to the POPATRISK data base.

Since the data base is county based, the most convenient and least expensive method of qualifying data is by use of C3 ST-COUNTY ID in a WHERE clause to qualify data sets for a particular county or span of counties, e.g. WHERE C3 EQ 343000. Component C3 is the only component in the data base that is unique for each logical entry.

The first component of each repeating group contains data which act as identifiers for each data set in the repeating group and are KEY'ed for use as selection criteria in WHERE clauses. For example, C101 NEDS POLLUTANT contains the five-digit criteria pollutant code KEY'ed for easy access to emissions data when used in the qualification of data, e.g., WHERE C101 EQ 11101. Each component name in the POPATRISK definition was chosen to be 16 characters or less to allow convenient default printout of the full component name in LIST Output.

The POPATRISK data base is structured such that all repeating groups, with the exception of SAROAD MONITORING SITES and MONITORING DATA are disjoint. Disjoint data sets are logically unrelated as shown in the data base structure in Figure 1. For retrieval purposes, it is most desirable to design a data base with repeating groups that are logically related whenever possible. However, each set of POPATRISK data, for the most part, came from a source for which there was no basis for a combination with other data types. For example, in-out migration age-groups do not correspond with age-specific age groups and NEDS emissions pollutant codes do not necessarily correspond with SAROAD monitoring pollutants. Therefore, data available for retrieval by means of simple S2K Natural Language commands, e.g., PRINT, LIST, are entry level data and data along any one branch of the tree structure in Figure 1.

Disjoint data sets may be retrieved using Natural Language in one of two ways. The user may use "BY clause" processing to normalize the retrieval process to a level of data that is common to families of all disjoint data sets to be retrieved. However, because of the nature of the POPATRISK data base, use of this retrieval method has not produced useful results. The other method of disjoint data set retrieval is to make retrievals without WHERE clauses. Since this method would qualify the entire data base, this method is not recommended. Therefore, using available Natural Language capabilities, separate retrievals should be made for sets of unrelated data. If further combining or processing of data is needed, each retrieval output may be written to a report file for input to a special utility program.

Also available to the user is the Report Writer feature of SYSTEM 2000. Although the feature requires learning a different set of retrieval techniques and commands, it provides the capability for preparing several reports within one run formatted specifically to the user's specifications. The merits of using Report Writer would be dependent entirely on its adaptability to future retrieval requirements.

Although it was not within the scope of this contract effort, an alternative solution to the limitations on retrievals of disjoint data sets discussed above is to develop either special purpose or generalized Procedural Language Interface programs to extract requested data from the data base. PLI offers the capability of establishing multiple positions within a data base entry to gain full access to all data sets and to structure output in a format most useful to the user. A PLI retrieval program would accept input parameter cards giving the type of data requested, specific items desired, selection criteria, and possibly data to be used in any calculations. Once the user's retrieval needs are assessed and PLI retrieval capability implemented, POPATRISK output could be provided in a timely and efficient manner and in a format tailored to the user's needs.

Another aspect of active data base use and maintenance that is not presently contemplated is updating of data. If, at some point, a decision is made to update existing data in the data base, PLI load programs could be developed to process new data and modify existing data and/or load new data sets or repeating groups. Any changes to the existing data base definition would in most cases require a reload of the data base.

SECTION 4

SAMPLE RETRIEVALS USING POPATRISK

In this section are presented a few sample retrievals intended primarily to demonstrate simple investigative capabilities of POPATRISK.

Table 5 illustrates one of the simplest types of retrievals available, a listing of all entries in a county for specific data items. In particular, the NEDS emission data, apportioned to all Massachusetts FIPS counties are listed. One special feature should be noted. The S2K Natural Language retrieval command which generated the listing of Table 5 calls for the item "*C1001*" which is a user-defined function that generates the data for area source emissions.

The result of another simple command is presented in Table 6. Here data on SAROAD monitoring for Durham County, N. C. are requested. There are three monitoring sites in the county; two in downtown Durham and one near a reservoir some distance to the north. Although the two downtown sites carry the same site code and are located at the same address, their UTM coordinates erroneously indicate actual locations some nine kilometers apart. Unfortunately coordinate errors of this type do occasionally occur in the SAROAD data base.

Table 7 lists a sample of counties in descending order of county air quality for Sulfur Dioxide averaged over sites in the county whose purpose was to monitor population exposure. The average for state-county code 192140, Orleans County, Louisiana, is unreasonably high and, most likely, the result of a keypunch error in the SAROAD data base. The potential user is cautioned that occasional errors do occur in SAROAD so that unusual data should be investigated thoroughly before significant conclusions are drawn. The SO_2 air quality does not appear to be closely correlated with any of the demographic or socio-economic data items chosen for display.

TABLE 5. ILLUSTRATIVE RETRIEVAL OF NEDS EMISSIONS DATA FOR MASSACHUSETTS IN TONS/YEAR

```
LIST/REPEAT SUPPRESS,TITLE(120)/C2,C101,C102,C103,
*C1001* WHERE C1 EQ 22:


*  COUNTY ID   NEDS POLLUTANT     COUNTY TOTAL     POINT SOURCES    AREA SOURCES
***
*   0001               11101          2519.500          1383.000        1136.500
*                      42401         33509.200         31682.000        1827.200
*                      42602         15257.100          8699.000        6558.100
*                      42101         57547.000          1202.000       56345.000
*   0003               11101          3169.600          1363.000        1806.600
*                      42401         14474.500         10727.000        3747.500
*                      42602         11085.300          2017.000        9068.300
*                      42101         70088.000          2192.000       67896.000
*   0005               11101         11437.100          6726.000        4711.100
*                      42401        112051.300        109956.000        2095.300
*                      42602         57436.700         36436.000       21000.700
*                      42101        141369.300          2858.000      138511.300
*   0007               11101           114.200            44.000          70.200
*                      42401           104.400             2.000         102.400
*                      42602           441.000            15.000         426.000
*                      42101          4012.200           227.000        3785.200
*   0009               11101          9085.900          2344.000        6741.900
*                      42401         30232.000         15274.000       14958.000
*                      42602         35539.400          3859.000       31680.400
*                      42101        224766.600          1128.000      223638.600
*   0011               11101          1080.300           500.000         580.300
*                      42401          2577.000          1297.000        1280.000
*                      42602          3017.400           328.000        2689.400
*                      42101         20403.200          1409.000       18994.200
*   0013               11101         31992.000         27003.000        4989.000
*                      42401         66652.600         55770.000       10882.600
*                      42602         46471.800         22517.000       23954.800
*                      42101        175591.600          3433.000      172158.600
```

(continued)

TABLE 5 (continued)

* 0015	11101	10364.600	9260.000	1104.600
*	42401	25213.900	22639.000	2574.900
*	42602	9961.800	4915.000	5046.800
*	42101	35815.500	1633.000	34182.500
* 0017	11101	19048.500	5176.000	13872.500
*	42401	42716.600	12562.000	30154.600
*	42602	69461.100	2860.000	66601.100
*	42101	490113.600	3248.000	486865.600
* 0019	11101	133.800	86.000	47.800
*	42401	98.600	5.000	93.600
*	42602	274.100	32.000	242.100
*	42101	2390.700	459.000	1931.700
* 0021	11101	4798.800	1080.000	3718.800
*	42401	16305.900	3905.000	12400.900
*	42602	28400.900	939.000	27461.900
*	42101	199605.100	750.000	198855.100
* 0023	11101	3300.900	173.000	3127.900
*	42401	7108.200	11.000	7097.200
*	42602	14559.600	64.000	14495.600
*	42101	101502.000	922.000	100580.000
* 0025	11101	9845.500	3325.000	6520.500
*	42401	21893.400	7324.000	14569.400
*	42602	30992.600	1658.000	29334.600
*	42101	204175.000	122.000	204053.000
* 0027	11101	20019.600	14437.000	5582.600
*	42401	34380.700	19081.000	15299.700
*	42602	39105.500	9228.000	29877.500
*	42101	219090.000	17810.000	201280.000

TABLE 6. ILLUSTRATIVE RETRIEVAL OF SAROAD SITE AND MONITORING DATA FOR DURHAM COUNTY, NC

```
PRINT/REPEAT SUPPRESS,NAME/C200 WHERE C301 EQ 11101
AND C3 EQ 341180:



   AREA CODE* 1160
   SITE CODE*    1
   AGENCY* G
   PURPOSE*  1
   CITY* DURHAM
   CITY POPULATION*    95438
   UTM ZONE* 17
   EASTING* 690551
   NORTHING* 3975943
   ADDRESS* HEALTH DEPT 300 EAST MAIN ST
   TYPE* 13
   ELEV ABOVE GND*   8
   ELEV ABOVE SEA*  405

     POLLUTANT* 11101
     METHOD* 91
     INTERVAL* 7
     # OBSERVATIONS*     42
     GEOMETRIC MEAN*     55.29
     STD DEVIATION*   1.96
     70TH PERCENTILE*    86.00
     90TH PERCENTILE*   115.00
     99TH PERCENTILE*   123.00
     HIGHEST VALUE*   123.00
     2ND HIGHEST VAL*   122.00
     LOWEST VALUE*     6.00

     POLLUTANT* 42401
     METHOD* 91
     INTERVAL* 7
     # OBSERVATIONS*     45
     GEOMETRIC MEAN*      6.55
     STD DEVIATION*   1.50
     70TH PERCENTILE*     8.00
     90TH PERCENTILE*    14.00
     99TH PERCENTILE*    17.00
     HIGHEST VALUE*    17.00
     2ND HIGHEST VAL*    16.00
     LOWEST VALUE*     5.00

     POLLUTANT* 42602                 70TH PERCENTILE*    40.00
     METHOD* 94                       90TH PERCENTILE*    50.00
     INTERVAL* 7                      99TH PERCENTILE*    71.00
     # OBSERVATIONS*     45           HIGHEST VALUE*    71.00
     GEOMETRIC MEAN*     21.44        2ND HIGHEST VAL*    69.00
     STD DEVIATION*   2.02            LOWEST VALUE*    10.00
```

(continued)

TABLE 6 (continued)

```
AREA CODE* 1160
SITE CODE*   1
AGENCY* P
PURPOSE*  1
CITY* DURHAM
CITY POPULATION*    95438
UTM ZONE* 17
EASTING* 689403
NORTHING* 3984981
ADDRESS* HEALTH DEPT 300 E MAIN ST
TYPE* 13
ELEV ABOVE GND*  50
ELEV ABOVE SEA*  405

  POLLUTANT* 11101
  METHOD* 91
  INTERVAL* 7
  # OBSERVATIONS*     6
  GEOMETRIC MEAN*    55.69
  STD DEVIATION*   1.21
  70TH PERCENTILE*    66.79
  90TH PERCENTILE*    69.49
  99TH PERCENTILE*    69.49
  HIGHEST VALUE*    69.49
  2ND HIGHEST VAL*    66.79
  LOWEST VALUE*    43.39

  POLLUTANT* 42401
  METHOD* 91
  INTERVAL* 7
  # OBSERVATIONS*     5
  GEOMETRIC MEAN*    15.35
  STD DEVIATION*   1.70
  70TH PERCENTILE*    22.19
  90TH PERCENTILE*    24.89
  99TH PERCENTILE*    24.89
  HIGHEST VALUE*    24.89
  2ND HIGHEST VAL*    22.19
  LOWEST VALUE*     7.09

  POLLUTANT* 42602
  METHOD* 94
  INTERVAL* 7
  # OBSERVATIONS*     6
  GEOMETRIC MEAN*    45.29
  STD DEVIATION*   1.10
  70TH PERCENTILE*    47.89
  90TH PERCENTILE*    49.79
  99TH PERCENTILE*    49.79
  HIGHEST VALUE*    49.79
  2ND HIGHEST VAL*    47.89
  LOWEST VALUE*    37.69
```

(continued)

TABLE 6 (continued)

```
AREA CODE* 1160
SITE CODE*   2
AGENCY* G
PURPOSE*  1
CITY* DURHAM
CITY POPULATION*     95438
UTM ZONE* 17
EASTING* 694973
NORTHING* 4002678
ADDRESS* LAKE MICHIE
ELEV ABOVE SEA*  403

  POLLUTANT* 11101
  METHOD* 91
  INTERVAL* 7
  # OBSERVATIONS*     24
  GEOMETRIC MEAN*     33.14
  STD DEVIATION*   1.67
  70TH PERCENTILE*     38.00
  90TH PERCENTILE*     59.00
  99TH PERCENTILE*     85.00
  HIGHEST VALUE*    85.00
  2ND HIGHEST VAL*     59.00
  LOWEST VALUE*     6.00

  POLLUTANT* 42401
  METHOD* 91
  INTERVAL* 7
  # OBSERVATIONS*     25
  GEOMETRIC MEAN*      5.21
  STD DEVIATION*   1.19
  70TH PERCENTILE*      5.00
  90TH PERCENTILE*      5.00
  99TH PERCENTILE*     12.00
  HIGHEST VALUE*    12.00
  2ND HIGHEST VAL*      6.00
  LOWEST VALUE*     5.00

  POLLUTANT* 42602
  METHOD* 94
  INTERVAL* 7
  # OBSERVATIONS*     25
  GEOMETRIC MEAN*     13.06
  STD DEVIATION*   1.33
  70TH PERCENTILE*     16.00
  90TH PERCENTILE*     17.00
  99TH PERCENTILE*     28.00
  HIGHEST VALUE*    28.00
  2ND HIGHEST VAL*     21.00
  LOWEST VALUE*    10.00
```

TABLE 7. ILLUSTRATIVE RETRIEVAL OF COUNTIES WITH HIGHEST CALCULATED MEAN SO_2 MONITORING (μ/m^3) FOR POPULATION MONITORING SITES INCLUDING SELECTED DEMOGRAPHIC DATA

```
LIST/TITLE(120)/C3,C10,C15,C16,C20,C402, ORDERED BY HIGH C402
WHERE C401 EQ 42401 AND C402 GT 30.00!


•  ST-COUNTY ID   TOTAL POPULATION   POP PER SQ MI   PERCENT URBAN   INCOME >15K/<5K   MEAN POP SITES
•••
•  192140               593471          3027.0            99.7               .47            1119.45
•  442740                11669            27.0              .0               .16             106.32
•  363160                96193           234.0            58.0               .73              89.30
•  181640                37370            80.0            17.8               .08              69.12
•  390100              1605016          2207.0            94.8              1.38              63.17
•  146900               285176           424.0            83.2               .86              62.96
•  511560               117917           435.0            71.4              1.40              59.66
•  334660              1539233         66923.0            99.9              1.12              56.80
•  361440               108310           203.0            55.9               .70              55.83
•  500200                29685           337.0            49.9               .71              55.49
•  470420                52637            57.0            36.7               .60              54.07
•  360540                80917           152.0            50.3               .45              53.00
•  264280               622236         10370.0            99.9               .49              50.90
•  397160              1948609         15223.0            99.9               .88              46.41
•  366700               232579           383.0            69.7              1.93              46.34
•  363820               303424           734.0            84.0              1.29              45.59
•  442800                35487            88.0            40.2               .19              43.27
•  330060               286742           545.0            85.7              2.01              42.88
•  366500               553371          1356.0            90.4              1.82              41.18
•  361600              1721300          3774.0            99.6              1.97              40.52
•  315440               543116          5324.0            99.9              3.72              39.68
•  501380                64197           611.0            86.2               .66              38.97
•  250680                35885            75.0            24.9               .25              38.87
•  244000                82948           215.0            69.0              2.79              38.60
•  312240               609266         12963.0            99.9              1.09              38.21
•  334720               235720           443.0            72.0              1.42              37.77
•  300300               223941           252.0            71.4              1.45              37.17
•  141440                28315            65.0            27.0               .53              36.19
•  313060               583813          1871.0            95.4              3.52              35.59
•  147600               118649           182.0            75.2              2.06              35.11
•  210120               905759         11763.0            99.9               .71              34.82
•  233620                39660            56.0            18.9               .49              34.44
•  140040                70861            82.0            63.9               .67              34.44
•  310740               456291          2074.0            95.9              1.88              34.37
•  391660               278311           365.0            45.0              3.07              34.22
•  410320               580261          1398.0            92.4               .94              34.07
•  513780               231365           418.0            80.2              4.51              33.78
•  264300               951353          1910.0            95.8              4.04              33.70
•  234620               120175           164.0            46.1              1.18              33.68
•  363720               484370          1412.0            94.1              1.66              32.71
•  070265               792814          1266.0            86.2              4.22              32.03
•  012400               317308           256.0            87.0               .43              31.51
•  141540              5492369          5759.0            99.7              2.28              31.31
•  491560               411027           245.0            82.5              1.12              30.64
•  399200               210876           246.0            44.0               .61              30.24
•  490340                34770            20.0            47.0               .92              30.04
```

The data in Table 8 were retrieved in order to investigate a possible relationship between county alcohol sales and the age-adjusted white male death rate for deaths attributable to motor-vehicle accidents. In order to place the numbers listed for alcohol sales/capita population over 18 in perspective, a feature of System 2000, namely the capability to produce nationwide averages for data items, was used in a separate retrieval to obtain the value of $64.40 for the nationwide average of alcohol sales > 18. Inspection of the data in Table 8 reveals that the counties reporting high death rates did not have unusually high per capita alcohol sales but that a disproportionate share of high death rate counties were rural counties. Data on gas sales/capita for State County Code 453330, Loving County, Texas, appear unreasonable and may be due to confidentiality restrictions in reporting of gas sales.

In Table 9 the age-adjusted death rates for white males for bronchitis are compared to those for influenza and pneumonia for counties having high death rates from bronchitis. A large proportion of the counties retrieved have small total populations and generally high migration rates. While the data suggest that there may be some relationship between the incidence of disease and population mobility, no conclusions should be drawn in this regard, nor with regard to possible association between the two disease categories, until more counties (especially urban counties) are included. The data base provides the capability of pursuing implications of high migration rates in counties with high (or low) death rates. This might be a logical next step since high migration rates for age groups not normally associated with the particular disease category under investigation would not, in and of themselves, indicate a county was unsuitable for drawing conclusions about possible associations.

The final retrievals in Tables 10 and 11 compare the geographical distribution of age-adjusted death rates for the white female population for breast cancer and for a cluster which includes breast cancer as a significant component. Eleven of the first 37 counties occur in both lists, but the data suggest a significantly different geographical distribution of the two variables. Similar retrievals of data using different variables could serve to illustrate the suitability of the cluster technique for investigating air quality-disease association.

TABLE 8. ILLUSTRATIVE RETRIEVAL OF COUNTIES WITH HIGHEST AGE-ADJUSTED AUTO FATALITIES FOR WHITE MALES AND ALSO SHOWING URBANIZATION, ALCOHOL, AND GAS SALES DESCRIPTORS

```
LIST/TITLE(120)/C3,C16,C22,C23,C703, ORDERED BY HIGH C703
WHERE C701 EQ 43 AND C703 GT 100.00:


*   ST-COUNTY ID    PERCENT URBAN    ALCOHOL SALE >18   GAS SALES/CAPITA   ADJ WMALE D-RATE
***
*   453330                     .0                47.7                 .0             496.00
*   011600                   26.3                21.2               12.1             322.34
*   452950                     .0                47.7              880.9             307.43
*   061060                     .0                76.1              288.3             305.00
*   454960                     .0                47.7              643.1             287.92
*   453250                     .0                47.7              242.4             251.21
*   060460                     .0                76.1              229.7             246.01
*   290300                     .0                75.4              579.1             230.02
*   113580                     .0                42.2              129.8             229.93
*   351060                     .0                51.8              406.9             229.22
*   290200                     .0                75.4              264.6             220.51
*   512120                     .0                43.7              451.5             219.70
*   271060                     .0                92.2              376.5             218.81
*   281300                     .0                36.0                 .0             216.98
*   450460                     .0                47.7              425.3             214.04
*   455180                     .0                47.7              328.2             213.88
*   451650                     .0                47.7               51.8             206.87
*   320660                     .0               102.4              416.1             203.94
*   455010                     .0                47.7              553.7             203.66
*   454340                     .0                47.7              223.9             203.12
*   184120                     .0                48.2              185.6             201.42
*   270920                     .0                92.2              130.6             200.28
*   453930                     .0                47.7              112.6             195.91
*   370640                     .0                44.8               78.9             190.75
*   060900                     .0                76.1               52.9             187.19
*   130340                     .0                41.6              251.4             186.75
*   114520                     .0                42.2               21.5             184.66
*   380860                     .0                87.9              206.0             184.26
```

(continued)

TABLE 8 (continued)

*	271600	.0	92.2	96.5	182.82
*	452870	.0	47.7	.0	181.70
*	060100	.0	76.1	322.4	181.46
*	054740	39.3	91.6	234.6	179.08
*	350080	.0	51.8	111.7	177.97
*	112340	.0	42.2	140.8	177.56
*	480760	.0	74.5	227.8	177.29
*	110700	35.8	42.2	317.7	177.04
*	111580	.0	42.2	123.9	175.70
*	450180	.0	47.7	480.3	175.11
*	280980	.0	36.0	268.7	174.34
*	061980	.0	76.1	420.7	173.20
*	320920	15.1	102.4	132.5	172.60
*	320180	.0	102.4	331.4	172.17
*	114860	.0	42.2	95.5	171.82
*	111940	.0	42.2	374.9	171.18
*	271220	47.1	92.2	150.4	169.82
*	270980	.0	92.2	281.7	167.88
*	111140	47.7	42.2	123.5	167.65
*	430600	.0	65.1	202.9	167.62
*	113740	34.0	42.2	429.8	166.67
*	170880	.0	58.8	144.9	163.75
*	181380	.0	48.2	186.3	161.94
*	010680	.0	21.2	124.2	160.03
*	113280	.0	42.2	112.7	159.98
*	112440	24.1	42.2	234.7	158.90
*	442700	.0	30.0	144.5	158.79
*	271680	.0	92.2	60.8	157.64
*	110320	.0	42.2	123.1	157.49
*	113380	.0	42.2	240.1	156.02
*	482100	.0	74.5	438.9	155.69
*	290460	.0	75.4	16.1	155.64
*	060160	.0	76.1	47.9	155.50
*	290520	.0	75.4	133.6	155.41
*	520360	62.6	96.4	543.4	153.31

(continued)

TABLE 8 (continued)

*	013460	.0	21.2	59.7	151.52
*	110360	38.9	42.2	151.3	150.99
*	454350	23.4	47.7	133.7	150.81
*	421280	18.4	54.1	200.8	150.30
*	451230	77.2	47.7	347.5	149.82
*	282380	.0	36.0	113.9	149.76
*	041720	50.9	51.6	173.7	149.75
*	252200	.0	32.0	95.6	149.16
*	350540	.0	51.8	125.5	149.10
*	441300	.0	30.0	79.6	149.09
*	061040	63.5	76.1	419.5	146.23
*	351100	.0	51.8	81.9	145.88
*	251180	21.2	32.0	131.0	145.22
*	062000	.0	76.1	161.3	145.02
*	172480	.0	58.8	338.6	144.90
*	110760	.0	42.2	81.4	144.17
*	280520	.0	36.0	211.6	144.00
*	290420	.0	75.4	507.2	143.45
*	271580	53.3	92.2	362.0	143.23
*	270620	.0	92.2	340.9	142.90
*	170760	.0	58.8	234.8	142.56
*	261080	26.5	49.3	199.3	141.99
*	251520	17.7	32.0	81.2	141.16
*	232780	.0	18.7	248.7	140.75
*	252580	13.6	32.0	84.7	140.46
*	271280	.0	92.2	561.4	139.63
*	371120	.0	44.8	187.4	138.81
*	101720	32.9	308.9	192.3	138.40
*	170580	.0	58.8	224.6	137.71
*	350240	.0	51.8	172.3	137.44
*	454450	36.0	47.7	130.2	137.43
*	454620	.0	47.7	103.8	136.93
*	183520	.0	48.2	36.0	136.83
*	321180	.0	102.4	176.0	136.80
*	114020	.0	42.2	11.5	136.73
*	431620	.0	65.1	153.0	136.12
*	112020	40.1	42.2	110.1	135.78

TABLE 9. ILLUSTRATIVE RETRIEVAL OF COUNTIES WITH HIGHEST WHITE MALE DEATH RATE FROM BRONCHITIS (CAUSE OF DEATH 28) ALONG WITH CORRESPONDING DEATH RATES FOR INFLUENZA AND PNEUMONIA (CAUSE OF DEATH 27), POPULATION, AND WHITE MALE MIGRATION DATA*

```
•  ST-COUNTY ID   TOTAL POPULATION   % WMALE IN MIG   % WMALE OUT MIG   CAUSE OF DEATH   ADJ WMALE D-RATE
•••
•  060900                     1272             45.6              48.5               27                .00
•                                                                                   28             180.94
•  271180                      675             18.7              59.9               27                .00
•                                                                                   28             179.33
•  350080                     1198              9.2              28.9               27                .00
•                                                                                   28             160.10
•  455670                     7344             27.8              38.9               27                .00
•                                                                                   28             143.47
•  455010                     3175             11.5              34.0               27              73.36
•                                                                                   28             139.98
•  110500                    10291             30.1              71.0               27              75.20
•                                                                                   28             137.72
•  114200                     8066             22.3              27.0               27              61.29
•                                                                                   28             136.36
•  290340                     2557             25.6              43.2               27              26.90
•                                                                                   28             129.23
•  171300                     5961             27.3              33.7               27                .00
•                                                                                   28             123.66
•  252660                    11854             20.8              28.2               27              19.07
•                                                                                   28             121.55
•  110780                    11334             20.0              22.5               27              96.72
•                                                                                   28             118.12
•  252940                    11099             15.4              20.4               27              38.92
•                                                                                   28             115.64
•  113200                     7044             41.7              24.3               27              14.05
•                                                                                   28             113.59
•  061120                     1811             29.4              45.8               27              92.11
•                                                                                   28             113.26
•  430520                     5170             17.4              29.5               27                .00
•                                                                                   28             105.49
•  460360                    12177             28.9              24.5               27              31.86
•                                                                                   28             105.47
•  351100                     1484             11.2              40.3               27                .00
•                                                                                   28             102.55
•  290520                      695             30.0              44.5               27             204.95
•                                                                                   28             102.48
•  110760                     6606             16.3              44.9               27             107.51
•                                                                                   28             102.00
•  111980                    17262              8.1              13.9               27              48.07
•                                                                                   28             100.80
•  113460                    12933              9.0              42.7               27              51.88
•                                                                                   28              96.64
•  420640                    29811             11.4              20.3               27              19.01
•                                                                                   28              95.88
```

---

* The above printout is the result of several combined retrievals to obtain the 40 counties with highest rates for bronchitis in descending order along with the corresponding rates for influenza.

(continued)

TABLE 9 (continued)

•	290180	629	31.1	52.7	27	93.57
•					28	93.57
•	062040	3405	21.3	33.9	27	66.73
•					28	93.34
•	101040	5480	18.6	28.9	27	38.43
•					28	92.12
•	060360	4819	46.1	28.1	27	48.47
•					28	91.97
•	453930	2258	37.5	30.5	27	68.76
•					28	91.57
•	010520	11824	19.3	24.0	27	104.82
•					28	90.46
•	321080	7189	30.5	18.6	27	36.21
•					28	88.88
•	053640	19548	36.6	22.9	27	43.86
•					28	87.33
•	060660	8407	43.8	20.4	27	39.99
•					28	87.11
•	060640	1641	24.4	55.3	27	56.63
•					28	86.19
•	114920	34515	19.2	22.0	27	63.95
•					28	86.06
•	110540	5940	16.4	20.4	27	30.74
•					28	85.47
•	130640	8710	24.3	27.0	27	26.79
•					28	85.31
•	110620	6539	23.5	27.3	27	64.75
•					28	85.21
•	470380	20153	12.2	14.5	27	57.91
•					28	84.54
•	061960	3827	16.8	38.4	27	51.86
•					28	83.88
•	454800	3657	29.5	34.7	27	32.12
•					28	83.53
•	450980	3087	17.0	32.2	27	17.08
•					28	83.19

TABLE 10. ILLUSTRATIVE RETRIEVAL OF COUNTIES WITH HIGHEST WHITE FEMALE AGE-ADJUSTED DEATH RATE FROM BREAST CANCER (CATEGORY 6) ALONG WITH POPULATION AND WHITE FEMALE MIGRATION DATA

```
LIST/TITLE(120)/C3,C10,C14,C44,C45,C704, ORDERED BY HIGH C704
WHERE C701 EQ 6 AND C704 GT 60.00:

•  ST-COUNTY ID   TOTAL POPULATION   PERCENT FEMALE   % WFEM IN MIG   % WFEM OUT MIG   ADJ WFEM D-RATE
•••
•  193160              11376              31.1             27.9            49.4              131.49
•  280200                847              49.6             13.5            56.4              131.07
•  130380                728              49.6              9.9            81.6              118.25
•  130460                741              47.4             25.9            62.6              118.25
•  114220               8394              50.3             17.9            13.5              106.50
•  431560               2457              50.0             22.8            37.0               98.57
•  281320               6807              50.2             22.6            20.0               96.55
•  011600              10650              53.6             16.3            29.2               90.32
•  271440               5779              49.1             10.9            26.1               88.23
•  461320               1483              48.2             20.5            36.7               87.70
•  061980                831              55.5             41.5            63.0               86.32
•  111440               5748              50.9             25.1            20.9               83.36
•  381900               1849              48.8             22.2            34.5               82.00
•  500400               6389              50.5             12.9            24.9               80.90
•  114660               6511              52.1             12.4            28.2               80.31
•  271180                675              43.0             23.1            58.5               76.31
•  290180                629              39.0             16.1            51.1               74.55
•  173480               2287              48.5             19.9            23.8               73.05
•  153180               4289              52.2             19.6            26.5               70.79
•  430220               1739              49.5             18.6            78.4               69.88
•  442440               3568              49.1             15.1            22.1               69.33
•  113980               7915              50.3             24.7            18.9               69.04
•  173280               5606              50.9             22.0            20.6               67.70
•  061240               8282              48.0             18.7            38.1               65.22
•  171820               3047              49.9             21.2            35.1               64.78
•  271020               2122              45.3             16.6            44.4               64.42
•  271200               5386              48.9             11.6            29.8               64.19
   271000               5014              48.7             30.8            23.0               64.12
•  421280              15878              52.0             13.8            19.1               63.99
•  061740               2185              50.3             49.0            31.7               63.52
•  454800               3657              49.2             27.7            35.1               63.28
•  171460               3672              49.3             27.3            25.8               63.22
•  251980              18983              51.8             18.6            18.5               62.95
•  265140               3359              51.8             11.3            25.2               62.85
•  241840              13351              49.2             16.5            24.8               60.92
•  393320               4926              50.0             26.3            25.9               60.40
•  113380               7371              50.7             19.8            23.1               60.03
```

TABLE 11. ILLUSTRATIVE RETRIEVAL OF COUNTIES WITH HIGHEST WHITE FEMALE AGE-ADJUSTED DEATH RATE FROM MALIGNANT NEOPLASM CLUSTER 1 (CATEGORY 49) ALONG WITH POPULATION AND WHITE FEMALE MIGRATION

```
LIST/TITLE(120)/C3,C10,C14,C44,C45,C704, ORDERED BY HIGH C704
WHERE C701 EQ 49 AND C704 GT 106.00:

•  ST-COUNTY ID   TOTAL POPULATION   PERCENT FEMALE   % WFEM IN MIG   % WFEM OUT MIG   ADJ WFEM D-RATE
•••
•  061980                831              55.5             41.5             63.0             229.45
•  431560               2457              50.0             22.8             37.0             203.88
•  271200               5386              48.9             11.6             29.8             157.44
•  234720              47392              51.2             17.9             17.7             157.02
•  192540               9937              51.8             16.3             16.3             153.33
•  282480                954              50.3             25.4             43.6             152.87
•  271360               6032              49.4             30.6             31.3             146.08
•  431840               2221              48.7             14.6             22.7             143.47
•  192700              18638              51.0             11.3             21.3             137.75
•  281320               6807              50.2             22.6             20.0             134.69
•  193160              11376              31.1             27.9             49.4             131.49
•  500400               6389              50.5             12.9             24.9             131.12
•  280200                847              49.6             13.5             56.4             131.07
•  171820               3047              49.9             21.2             35.1             129.61
•  481580              18285              50.7             17.8             17.9             129.43
•  171300               5961              49.7             24.7             28.6             127.36
•  510960               3298              48.4             15.8             30.2             123.80
•  192620              32453              51.1              9.0             11.1             121.83
•  130460                741              47.4             25.9             62.6             118.25
•  130380                728              49.6              9.9             81.6             118.25
•  442440               3568              49.1             15.1             22.1             117.67
•  452590               2392              50.8             28.5             73.0             116.53
•  130040               2877              48.3             31.8             44.8             115.73
•  290560              10150              49.5             23.6             25.1             115.27
•  113380               7371              50.7             19.8             23.1             114.63
•  271440               5779              49.1             10.9             26.1             113.75
•  111940               1924              50.6             19.7             17.8             110.16
•  430240               7825              50.8             18.8             30.5             109.93
•  280320               4021              51.4             21.9             37.0             109.03
•  282060               6468              51.5             11.7             24.5             108.82
•  192520              29550              49.9             19.8             16.0             108.30
•  140540              25440              50.2             26.7             19.8             107.59
•  265140               3359              51.8             11.3             25.2             106.76
•  114220               8394              50.3             17.9             13.5             106.50
•  180860               6850              51.1              8.6             15.6             106.47
•  270540               1796              52.1             10.4             26.4             106.30
•  350160               3901              49.5             14.3             33.8             106.06
```

SECTION 5

DATA BASE DEVELOPMENT PROCEDURES

CONSIDERATIONS LEADING TO CHOICE OF DATA ITEMS

Choices of specific data items to be included in POPATRISK were made in close consultation with the EPA Project Officer. Considerations of utility to EPA were primary but the limitations imposed by data base size and consequent development and retrieval times played a significant role in the decision process. In this section, some rationale is given for the specific choices made.

Entry Level Data Items

The county level is the smallest geographic entity for which deaths and population mobility data are available. In addition, area emission estimates provided by the NEDS file are available only on a county basis. Even if these data sources were not restricted to the county level, daily population mobility such as commuting to work, strongly suggests that a there is not much to be gained in studies of possible relationships between air quality and disease in a finer geographic subdivision than the county unless data on individuals themselves are presented. For these reasons, the county was chosen as the geographical entity for POPATRISK. Once this decision was made, it was decided to include data on population density, racial makeup, measures of socio-economic status, climatology, and population mobility both with regard to work/residence and to in- and out-migration.

Monitoring Data

It was decided to preserve in its entirety the integrity of monitoring data in SAROAD. The inclusion of UTM coordinates, e.g. makes feasible the development of air quality estimates on a finer than county level for special user purposes. The geometric mean and standard deviation enable a ready computation of arithmetic mean, if the user wishes, and might even be accomplished by a user-defined function if experience shows this statistical measure is in much demand.

Air Quality

A number of schemes of varying complexity were considered as candidates for the estimation of county air quality. In the end, the simplest, i.e., an arithmetic mean of geometric means for all stations by purpose type, was chosen because no other seemed to offer distinct advantages. As mentioned above both intraarea and interarea population mobility and the coarseness of mortality statistics were considered to limit the usefullness of a finer geographical breakdown. It was felt also that in preserving the SAROAD monitoring data intact, POPATRISK would enable individual investigators to utilize the data for constructing more complex air quality assessments to meet particular analytical needs with maximum flexibility.

In-out Migration

The Census tape from which POPATRISK data on in-out migration were derived, was specially compiled by Census for this work and represents to our knowledge, the first use of these data at the county level. It was felt especially important to include these data so that individual assessment could be made of the stability of a county with regard to population and consequently so that possible artifacts in the relationship between air quality and mortality could be flagged and, perhaps, explained.

Mortality

The direct method of age-adjustment was chosen to correspond to the practice most widely used for making county mortality statistics comparable from county to county. Limitations on data base size precluded more than about 50 ICDA categories so that it was decided to choose the 48 categories corresponding to aggregations of ICDA's already in use at HERL. Following a published attempt in the literature to provide the most rational basis for aggregating ICDA's into meaningful groups it was decided to include two other aggregations on an experimental basis. These aggregations were derived from the results of a cluster analysis reported by F. Burbank (A Sequential Space-time Cluster Analysis of Cancer Mortality in the United States: Etiologic Implications, Am. J. Epid 95, 383-417, 1972) in which ICDA cancer mortalities were analyzed on the basis of geographical similarity among tumor types, separately for the white male and white female population. For this work, it was decided to select groupings of ICDA's which were quite similar for the white male and white female population. With this selection criterion only two clusters seemed meaningful and were used for the remaining two mortality categories. The utility of this approach remains a subject for future investigation.

HISTORY OF DEVELOPMENT

The POPATRISK data base development and loading procedure was modularized, in that, as data from each selected source became available, methodologies for processing and loading the data along with resolution of geocoding discrepancies were developed and implemented.

Initial verification of raw data to be used in the data base involved for the most part close inspection of the data and extensive correspondence with those persons responsible for providing the data. Raw data verification processes were performed on Census and mortality data during processing procedures and will be discussed later in this section.

A more extensive verification process was performed on the in-out migration data obtained from the Bureau of the Census. The verification effort was done after consultation with the Bureau of the Census because its use in POPATRISK represented the first access to this data outside the Bureau. Detailed documentation on the migration file is provided in Appendix B.

To verify the data on the in-out migration file, 33 counties which are also state economic areas were selected and data for these counties were compared with published Census data. The number of unallocated out-migrants by sex and by age group in these counties matched exactly the data in Table 2 of the PC(2)-2E Migration Between State Economic Areas Census report.

The numbers of unallocated in-migrants on the file were also checked for these counties. Since comparisons with published reports could not be made by sex or by age group, in-migrant totals on the file were summed for both sexes and all ages. These unallocated in-migrant totals on the file were compared with totals obtained by adding the "abroad in 1965" data in Table 1 of the PC(2)-2E Migration Between State Economic Areas to the sum of the male and female "population 5 years old and over" data in Table 2 of the PC(2)-2E publication. Exact matches could not be made with published data on in-migration because of differing allocation procedures used in creating the migration file, but the differences in the data totals of each county selected were less than 1%. Table 12 gives the results of this comparison.

State totals on the tape were also verified for the contiguous United States and the District of Columbia by aggregating totals for all ages and both sexes. These totals corresponded exactly with the unallocated out-migrant totals determined from Table 4 of PC(2)-2E by the following formula: "total persons living in state in 1965" - "total persons still living in state in 1970."

As with the in-migrant county totals, unallocated in-migrant state totals on our file differed slightly (less than 1%) from published data. Data from Table 1 of PC(2)-2E under the general heading "different state economic area" and the specific heading "different state" were added to "abroad in 1965" data from Table 44 of PC(2)-2B, Mobility for States and the Nation. The resulting data totals compared closely with the sum of male and female in-migrant totals for all ages on our tape. This comparison is shown in Table 13.

METHODOLOGIES FOR PROCESSING AND LOADING

During the entire procedure of processing and loading the POPATRISK data base, extreme care was taken to assure proper maintenance of the quality and integrity of the data from their verified sources, through the processing stage, and into the data base. This involved a thorough investigation of the data to isolate all anomalies and geocoding discrepancies in the data

TABLE 12. UNALLOCATED IN-MIGRATION TOTALS (ALL RACES, AGES, BOTH SEXES) FOR SELECTED COUNTIES WHERE COUNTY=STATE ECONOMIC AREA (SEA)

STATE	COUNTY	FIPS CODE	SEA	TOTAL 1*	TOTAL 2**
Arizona	Pima	04 019	B	95501	95371
Arkansas	Pulaski	05 119	A	52092	52129
California	Ventura	06 111	7	114038	113820
Colorado	El Paso	08 041	B	94668	94312
Connecticut	Hartford	09 003	C	95197	94403
Delaware	New Castle	10 003	A	52501	52338
Georgia	Bibb	13 021	F	16985	16988
Illinois	Boone	17 007	2	6545	6568
Indiana	St. Joseph	18 141	B	29369	29379
Iowa	Woodbury	19 193	A	14571	14584
Kansas	Sedgwick	20 173	A	62292	62222
Kentucky	Fayette	21 067	E	38991	38985
Louisiana	Calcasieu	22 019	D	18725	18722
Maine	Aroostook	23 003	1	13273	13154
Massachusetts	Worcester	25 027	B	61123	60888
Michigan	Kent	26 081	B	49305	49343
Mississippi	Hinds	28 049	A	31879	31881
Missouri	Greene	29 077	C	33160	33181
Nevada	Clark	32 003	A	73634	73572
New Hampshire	Hillsborough	33 011	A	35751	35796
New Jersey	Atlantic	34 001	E	21378	21395
New Mexico	Bernalillo	35 001	A	68623	68495
New York	Broome	36 007	E	25542	25542
North Carolina	Wake	37 183	E	51031	51005
Ohio	Allen	39 003	O	12341	12330
Oklahoma	Creek	40 037	C	9506	9506
Oregon	Lane	41 039	B	50693	50676
Pennsylvania	Blair	42 013	F	9868	9868
South Carolina	Aiken	45 003	B	13024	12997
Tennessee	Davidson	47 037	B	66472	66346
Utah	Salt Lake	49 035	A	57067	56803
Washington	Spokane	53 063	D	58094	58011
West Virginia	Kanawha	54 039	C	22999	22861

* Total from 1970 migration tape
** Total from Census publication

TABLE 13. STATE UNALLOCATED IN-MIGRATION TOTALS FOR ALL RACES, AGES, BOTH SEXES

STATE	FIPS CODE	TOTAL FROM 1970 MIGRATION TAPE	TOTAL FROM CENSUS PUBLICATION
Alabama	01	253274	253445
Arizona	04	367799	368192
Arkansas	05	183763	183918
California	06	2164948	2171140
Colorado	08	393295	393764
Connecticut	09	310808	311115
Delaware	10	74058	74175
District of Columbia	11	100313	100436
Florida	12	1344272	1345594
Georgia	13	474952	475353
Idaho	16	100797	100926
Illinois	17	797673	798727
Indiana	18	387630	388114
Iowa	19	189704	190120
Kansas	20	261540	261819
Kentucky	21	239756	239915
Louisiana	22	248315	248684
Maine	23	82386	82527
Maryland	24	513130	513893
Massachusetts	25	450030	450842
Michigan	26	564476	565340
Minnesota	27	269816	270656
Mississippi	28	168462	168549
Missouri	29	410022	410605
Montana	30	76863	76972
Nebraska	31	138803	138998
Nevada	32	119133	119224
New Hampshire	33	106157	106285
New Jersey	34	723874	724639
New Mexico	35	144420	144617
New York	36	1053286	1054876
North Carolina	37	415305	415688
North Dakota	38	59260	59322
Ohio	39	681849	683030
Oklahoma	40	303150	303425
Oregon	41	273894	274395
Pennsylvania	42	582853	584037
Rhode Island	44	103951	104311
South Carolina	45	246710	246959
South Dakota	46	56105	56177
Tennessee	47	326387	326679
Texas	48	1028483	1029826
Utah	49	118723	118865
Vermont	50	56911	56940
Virginia	51	641962	643407
Washington	53	512079	513084
West Virginia	54	114621	114722
Wisconsin	55	270043	270534
Wyoming	56	52575	52611

as well as verification that all such problems were resolved as accurately as possible. The effort was accomplished, for the most part, through use of county code cross-walk files developed to match record identifiers with data base entries. These files were merged to produce the county cross-reference table included in the POPATRISK Geocoding Manual.

Every opportunity was taken to spot-check data during processing and loading to insure not only accuracy of processing but also that data were loaded into the proper county entry and in the correct location. The cohesive nature of the data base entries provided opportunity for an on-going process of verifying already loaded data as new data were being processed and loaded.

Figure 2 provides an overview of the first phase of the POPATRISK data base development process. The loading processes shown in Figure 2 were fairly mechanical in nature and involved relatively moderate amounts of data. For these sets of data the loading procedure involved extracting data from the input files and structuring the data into S2K load string format with the proper associated county code and then performing Natural Language Queue Access loading. All loading processes involved incremental loads in groups of county repeating groups to assure successful execution of load jobs within the constraints of run time and file sizes.

Population demographics and climate data items in the entry level were taken straight from the SSI-developed tape and loaded county by county. Matching of counties in the data base presented no problems with this data since discrepancies had been resolved in previous work by SSI. However, as mentioned in Section 2, entry level data for New York City were provided in one record for New York City as a whole. Income ratios and climate data were considered the same for the five boroughs and were duplicated for each borough entry. Other demographic data for each borough were extracted from the 1972 County and City Data Book.

The NEDS emissions and SAROAD monitoring data were also simply extracted from the appropriate NADB files and loaded into the data base county by county. Users should be aware that some monitoring data in

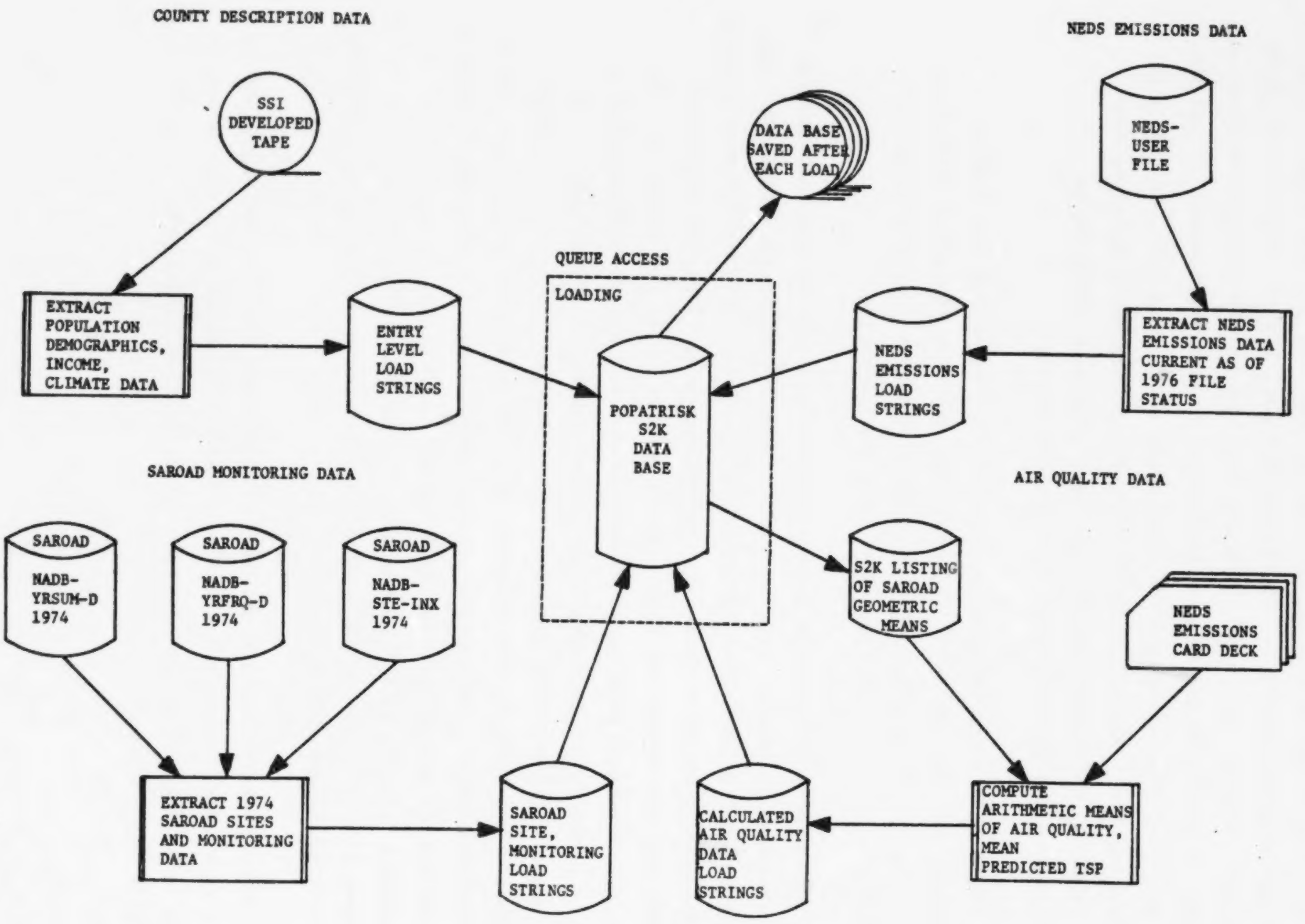

COUNTY DESCRIPTION DATA
SSI DEVELOPED TAPE
EXTRACT POPULATION DEMOGRAPHICS, INCOME, CLIMATE DATA
ENTRY LEVEL LOAD STRINGS
QUEUE ACCESS
LOADING
POPATRISK S2K DATA BASE
DATA BASE SAVED AFTER EACH LOAD
NEDS EMISSIONS DATA
NEDS-USER FILE
EXTRACT NEDS EMISSIONS DATA CURRENT AS OF 1976 FILE STATUS
NEDS EMISSIONS LOAD STRINGS
SAROAD MONITORING DATA
SAROAD
NADB-YRSUM-D 1974
SAROAD
NADB-YRFRQ-D 1974
SAROAD
NADB-STE-INX 1974
EXTRACT 1974 SAROAD SITES AND MONITORING DATA
SAROAD SITE, MONITORING LOAD STRINGS
CALCULATED AIR QUALITY DATA LOAD STRINGS
AIR QUALITY DATA
S2K LISTING OF SAROAD GEOMETRIC MEANS
NEDS EMISSIONS CARD DECK
COMPUTE ARITHMETIC MEANS OF AIR QUALITY, MEAN PREDICTED TSP

Figure 2. Overview of POPATRISK data sources and development procedures, Phase 1.

SAROAD is of questionable nature. This applies for the most part to a few of the reported values for observed maxima. It was not possible, within time and budget constraints, to edit SAROAD data for the purposes of this contract. However, it is our judgment that the yearly averages abstracted for use in POPATRISK can be used with considerable confidence. It should also be noted that although running-average monitoring data, identified by sampling intervals X, Y, or Z, were initially loaded into the data base, quality and the usefulness of these data were found to be questionable and the data were subsequently removed.

Since NEDS recognizes only 6 counties in Massachusetts, SAROAD recognizes 13, and the POPATRISK data base contains 14 Census counties, resolution of the discrepancies required a complete reapportionment of both sets of data to the respective data base counties.

As discussed in Section 2, the Census coding scheme was selected for Massachusetts because of its finer geographic breakdown and widespread use. The four-digit county codes used are the three-digit FIPS codes with leading zeros added to create a four-digit POPATRISK code. Table 14 lists each SAROAD city and county and POPATRISK code to which it was assigned. The NEDS area emissions were aggregated for the state and then disaggregated into 14 Census counties using 1970 Census figures, gasoline sales, etc., on a basis which conforms largely to the guidelines in the EPA document 'Guide for Compiling a Comprehensive Emission Inventory' APTD-1135. There are approximately 2,500 point sources in Massachusetts. They were assigned to appropriate Census counties by a computerized procedure involving geographic location by UTM coordinates. A detailed discussion of the procedures used to assign NEDS point source data to Massachusetts counties is provided in Appendix A.

SAROAD monitoring data for Massachusetts were simply reassigned to their appropriate data base counties by identifying the city name of all SAROAD monitoring sites and then locating the data base county in which the city is located.

TABLE 14. ASSIGNMENT OF MASSACHUSETTS SAROAD MONITORING TO CENSUS COUNTIES (all state 22)

SAROAD City Name	Census County Name	POPATRISK County Code
Adams	Berkshire	0003
Amherst	Hampshire	0015
Athol	Worcester	0027
Attleboro	Bristol	0005
Ayer	Middlesex	0017
Belchertown	Hampshire	0015
Boston	Suffolk	0025
Brookline	Norfolk	0021
Cambridge	Middlesex	0017
Chicopee	Hampden	0013
Fall River	Bristol	0005
Fallmouth	Barnstable	0001
Fitchburg	Worcester	0027
Framingham	Middlesex	0017
Greenfield	Franklin	0011
Haverhill	Essex	0009
Holyoke	Hampden	0013
Lawrence	Essex	0009
Lee	Berkshire	0003
Lowell	Middlesex	0017
Lynn	Essex	0009
Marblehead	Essex	0009
Maynard	Middlesex	0017
Medford	Middlesex	0017
Needham	Norfolk	0021
New Bedford	Bristol	0005
Newburyport	Essex	0009
North Adams	Berkshire	0003
Northfield	Franklin	0011
Norwood	Norfolk	0021
Peabody	Essex	0009
Pittsfield	Berkshire	0003
Plymouth	Plymouth	0023
Quincy	Norfolk	0021
Revere	Suffolk	0025
Springfield	Hampden	0013
Waltham	Middlesex	0017
Warren	Worcester	0027
Woburn	Middlesex	0017
Worchester	Worcester	0027

Since POPATRISK is a county based data base, data for independent cities without county status required reassignment to their original counties. Only three independent cities are universally given county status and therefore maintain that status in the data base. They are Baltimore, Maryland, St. Louis, Missouri, and Carson City, Nevada and are listed in Table 15. Independent cities in Virginia were assigned to their original counties as shown in Table 16. As mentioned in Section 2, it was necessary to create two data base counties for those independent cities for which no county now exists, namely Elizabeth City - Warwick and Norfolk - Princess Anne, with arbitrarily assigned codes 0972 and 2144 respectively. Entry level data for the new counties were obtained by aggregating data for independent cities contained therein.

The NEDS emission data for Virginia independent cities were added to the NEDS data for the county to which they were assigned. Independent city SAROAD site and monitoring data were simply retrieved and loaded into the county entry to which they were assigned.

To calculate monitoring air quality data, the geometric means for each pollutant were retrieved from the data base for purposes of (1) population monitoring, (2) source monitoring, and (3) background monitoring where the sampling interval is 1 (1 hour) or 7 (24 hours). Arithmetic averages of the geometric means were calculated for each pollutant for each of the three site types.

A predicted mean of emissions for TSP was calculated by SSI on the basis of a regression model described in Appendix C. A card deck of NEDS data had to be used because the regression model was developed using a version of the NEDS emissions inventory which is no longer currently accessible through NADB files. The predicted mean for TSP was calculated for counties where TSP and NOx emissions (tons/year) fall within the following limits:

$$1414.375 \leq \text{TSP} \leq 230031.803$$
$$\text{NOX} \leq 432778.195$$

TABLE 15. INDEPENDENT CITIES RECOGNIZED AND CODED IN DATA BASE

City Name, State Name	POPATRISK State-County Code	
Baltimore, Maryland	21	0120
St. Louis, Missouri	26	4280
Carson City, Nevada	29	0040

TABLE 16. ASSIGNMENT OF VIRGINIA INDEPENDENT CITIES TO POPATRISK COUNTIES

(all state 48)

City Name	SAROAD Code	County Name	POPATRISK County Code
Alexandria	0080	Arlington	0200
Bedford	0320	Bedford	0340
Bristol	0480	Washington	3300
Buena Vista	0560	Rockbridge	2740
Lexington	1740	Rockbridge	2740
Charlottesville	0680	Albemarle	0060
Clifton Forge	0780	Alleghany	0100
Covington	0840	Alleghany	0100
Colonial Heights	0820	Chesterfield	0720
Danville	0920	Pittsylvania	2380
Emporia	0980	Greensville	1400
Fairfax	1040	Fairfax	1060
Falls Church	1080	Fairfax	1060
Franklin	1180	Southampton	2940
Fredericksburg	1240	Spotsylvania	3000
Galax	1280	Grayson	1360
Harrisonburg	1480	Rockingham	2760
Hopewell	1560	Prince George	2500
Lynchburg	1840	Campbell	0580
Martinsville	1940	Henry	1520
Norton	2240	Wise	3420
Petersburg	2360	Dinwiddle	0960
Radford	2600	Montgomery	2020
Richmond	2660	Henrico	1500
Roanoke	2700	Roanoke	2720
Salem	2800	Roanoke	2720
South Boston	2920	Halifax	1420
Staunton	3060	Augusta	0260
Waynesboro	3320	Augusta	0260
Suffolk	3080	Nansemond	2060
Williamsburg	3360	James City	1600
Winchester	3380	Frederick	1220
Hampton	1440	Elizabeth City-Warwick	0972
Newport-News	2120	Elizabeth City-Warwick	0972
Chesapeake	0710	Norfolk-Princess Anne	2144
Norfolk	2140	Norfolk-Princess Anne	2144
Portsmouth	2440	Norfolk-Princess Anne	2144
Virginia Beach	3240	Norfolk-Princess Anne	2144

A formatted input file of SAROAD data selected from the data base was read simultaneously with NEDS input card deck. The county identifiers on each file were matched and calculations performed. Results were placed in S2K load strings and subsequently loaded into the data base.

In those cases for which there were no monitoring data for a county, but the TSP and NO_x values qualified the county for calculation of mean predicted TSP, a single data set for TSP exists containing the mean predicted TSP.

The air quality repeating group provides the user readily accessible estimates for assessing the general exposure of county population to the various pollutants.

Figure 3 provides an overview of the second phase of the data base development process. Processing and loading of data at this stage involved numerous calculations and handling of much larger quantities of data. Therefore, S2K Procedural Language Interface COBOL programs were used to interface with the data base, retrieve data base data needed for calculations, and load the data by means of PLI optimized loading.

The Census in-out migration tape was read and the FIPS county code for each record on the tape was written to a file to create a cross-walk file for matching migration counties with data base counties. Although there are state total records on the original tape, these data were not needed for POPATRISK and were disregarded. Also disregarded on the tape were migration records for Alaska and Hawaii.

Geocoding problems with the migration data were resolved in a manner similar to that used for previous data types. Baltimore and St. Louis independent cities and counties were reordered to match the SAROAD scheme; data for Yellowstone Park, Montana was added into the record for Park

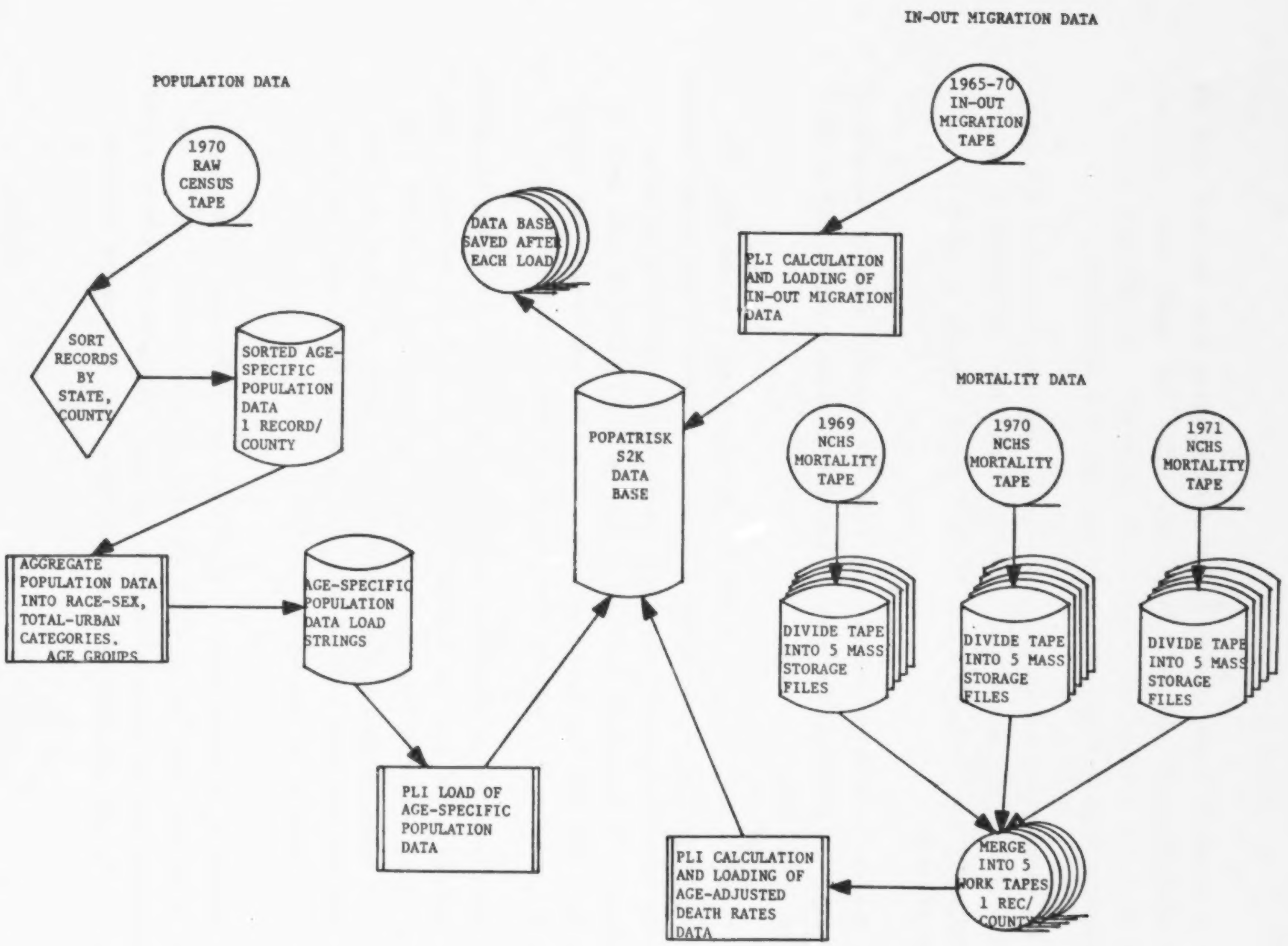

IN-OUT MIGRATION DATA
POPULATION DATA
1970 RAW CENSUS TAPE
SORT RECORDS BY STATE, COUNTY
SORTED AGE-SPECIFIC POPULATION DATA 1 RECORD/ COUNTY
AGGREGATE POPULATION DATA INTO RACE-SEX, TOTAL-URBAN CATEGORIES, AGE GROUPS
AGE-SPECIFIC POPULATION DATA LOAD STRINGS
PLI LOAD OF AGE-SPECIFIC POPULATION DATA
DATA BASE SAVED AFTER EACH LOAD
POPATRISK S2K DATA BASE
1965-70 IN-OUT MIGRATION TAPE
PLI CALCULATION AND LOADING OF IN-OUT MIGRATION DATA
MORTALITY DATA
1969 NCHS MORTALITY TAPE
1970 NCHS MORTALITY TAPE
1971 NCHS MORTALITY TAPE
DIVIDE TAPE INTO 5 MASS STORAGE FILES
DIVIDE TAPE INTO 5 MASS STORAGE FILES
DIVIDE TAPE INTO 5 MASS STORAGE FILES
MERGE INTO 5 WORK TAPES 1 REC/ COUNTY
PLI CALCULATION AND LOADING OF AGE-ADJUSTED DEATH RATES DATA

Figure 3. Overview of POPATRISK data sources and development procedures, Phase 2.

County, Montana, and data for all Virginia independent cities were added into records for their corresponding counties. There was no record on the migration tape for Loving, Texas, POPATRISK code 45 3330. Since the population of that county was 164, the data would have been suppressed by Census confidentiality requirements.

When all geocoding problems were resolved, a PLI program was written to process the migration tape using the FIPS - POPATRISK cross-walk file and load the data county by county. Processing of migration data involved retrieval of the 16th data set (totals) of the age-specific Census data. These race-sex totals were used to convert migration data to percentages of the total population for a particular race-sex group. For example, the number of white male in-migrants for age group 14 was divided by total white males in the county and multiplied by 100 to give a percentage of the total population of white males that were both in-migrants and between age 5 and 14. Migration percentages were calculated for both 'Total' (Allocated + Unallocated) and 'Allocated' data. As data sets were processed and loaded, total county migration figures were tallied and loaded into the entry level components for the county.

The New York City boroughs presented a unique problem with in-out migration data. As the Census documentation in Appendix B explains, persons reporting in 1970 that their residence in 1965 was merely a different house in New York City or an incomplete response of "New York City - no borough specified" were not included in either the in- or out-migrant tallies for individual boroughs. An extra summary record for New York City as a whole is provided that excludes those persons with the above responses. The documentation also describes another extra record on the tape that contains out-migrant counts for those persons with a 1965 residence of "New York City - no borough specified" and 1970 residence outside New York City. However, this record was not found on the tape.

In order to complete the data for New York City boroughs, SSI contacted the Bureau of the Census to discuss the possibility of retrieving data not included on the tape and incorporating that data into the data base. The Census Bureau was very helpful in providing hard copy tabulations of the needed data. Specifically, tabulations provided show counts of persons reporting 1970 residence of one of the New York City boroughs and 1965 residence of 'New York City' - borough not specified by sex, race, and age groups; also supplied were counts of persons reporting 1970 residence outside New York City and 1965 residence 'New York City' - no borough specified by race, sex, and age group.

A procedure was developed to distribute among boroughs the count of in-migrants who gave incomplete responses (no borough specified) to 1965 residences, between the five boroughs according to a ratio proportional to in-migrants giving complete responses. The procedure resulted in estimated counts of 'Incomplete response' in-migrants from each of the five boroughs, which when summed over the 1970 residence boroughs, provided out-migrant counts from each of the boroughs. Both in- and out-migrant figures were considered 'allocated' and were added to their respective borough records.

'Incomplete response' out-migrants from New York City were distributed to each of the five boroughs according to the 1970 borough population. These figures were also considered 'allocated' and where added to their appropriate boroughs.

The procedure for processing the NCHS mortality data and calculating age-adjusted rates was as follows:

To ease the problem of working with such large numbers of records on the raw tapes, each tape was divided into five separate mass storage files. Since the raw data were sorted by state and county, three mass storage files, one from each tape, could be read and manipulated simultaneously resulting in a merge of all deaths over the three-year period for each county into a worktape record. Counts of deaths were stored in the worktape record in a three dimensional age(15), race-sex(4), ICDA(50) array. The 15 age groups used for mortality data are identical to those used in the age-specific population data. Five runs were made, using three mass storage raw data files at a time, to create five separate tapes of mortality counts, one record per county. These five work tapes were then saved to be combined into one file spanning two tapes for delivery to EPA along with this final report. Documentation on the work tapes are provided with the tapes as well as in Appendix B of this report.

Direct method age-adjusting was accomplished as follows:

Let the indices i, j, k be defined as:

i	ICDA (1-50)
j	Age Group (1-15)
k	Race-sex Group (WM, WF, NWM, NWF)

and let:

$Dijk$ = total number dead in county for years 1969, 1970, 1971 in ICDA i cause category, age group j, and race-sex group k, taken from the work tape record.

Rjk = number of people in age group j and race-sex group k in county (i.e. number at risk)

$Sijk = \frac{Dijk}{Rij}$ = age-specific death rate in ICDA i cause in age group j and race-sex group k

Pj = number of people in age group j in the standard U.S. population in base year 1970.

$P = \sum_{j} Pj$ = total U.S. population in 1970

Aik = age-adjusted death rate in ICDA i cause in race-sex group k (per 100,000)

Aik can be written as:

$$Aik = \frac{\sum_{j=1}^{15} Sijk\ Pj}{P}$$

with a substitution for Sijk:

$$Aik = \sum_{j=1}^{15} \frac{Dijk}{Rjk} \frac{Pj}{P}$$

The actual calculation of age-adjusted rate per 100,000 over 3 year period is performed as follows:

$$\overline{Dijk} = \frac{Dijk}{3} = \text{average number of deaths over 3 year period}$$

$$Aik = \sum_{j=1}^{15} \frac{\overline{Dijk}}{Rjk} \cdot \frac{Pj}{P} \cdot 100{,}000$$

The variable R_{jk} is easily retrieved from the AGE-SPECIFIC POPULATION repeating group. Since the number dead, Dijk, is divided by three, the death rates represent the average yearly age-adjusted death rate per 100,000 population over the three-year period from 1969 to 1971.

Verification of the mortality data and processing methodology began with the mortality count worktape records prior to any calculations and loading. Three test files were created for testing the worktape creation software. When verified correct, the worktapes were created and extensive spot-checking of the worktapes was accomplished using published NCHS publications. Data is published for each year, showing total deaths for each ICDA category for each county. Adding deaths across age groups and across the three years provided verification of total death counts on the worktapes.

Geocoding problems with the NCHS coding scheme were handled in a similar manner to the in-out migration data. Mortality worktape records containing counts for Virginia independent cities were combined with the worktape records for their respective counties. With a few exceptions, the worktape delivered to EPA contains all geocoding corrections giving a county by county match-up of worktape records with data base counties. Mortality data were available only for New York City as a whole with no breakdown for the five boroughs. Therefore, only one record exists on the worktape for the five New York City boroughs.

A PLI COBOL program was developed to extract death counts from the worktape, retrieve county Census data, calculate age-adjusted rates and load the rates into the data base, using PLI optimized load. Upon completion of testing of the mortality calculations and loading process and trial loading of data for several counties, it became apparent that processing time and charges for loading the data in their present format would be prohibitive. Version 2.80 of SYSTEM 2000 offers the capability for creation and removal of indices at any time, in effect, changing components from KEY to NON-KEY, and vice versa. Therefore, all the KEY'ed components in the mortality repeating group were charged to NON-KEY during the loading process. This action eliminated the excessive processing time and expense of creation and maintenance of KEY item pointers in the data base files.

A decision was made to process and load the mortality data for New York City into each of the five New York City boroughs using the total population of the five boroughs for calculations. This resulted in identical data for each of the boroughs which represent age-adjusted death rates for the whole of New York City.

At the present time, there are no published data directly comparable to the direct method age-adjusted death rates present in the data base. Although NCHS did provide unpublished age-adjusted data for 1969-71, they customarily use the 1940 standard million for calculations. Therefore, to verify the data, extensive hand calculations were done for randomly selected counties. In addition, calculations were done using the 1940 standard million to arrive at results comparable to NCHS data. Calculated data corresponded extremely well with the NCHS data. Any slight variances in results could be explained by the slight differences in the distribution of 1970 county population data across race groups.

It should be noted that in some cases, calculated age-adjusted death rates may be misleading, specifically in cases where there are few or no people in a race-sex-age group in a county. The situation occurs, for example, in counties where there are few, if any, non-white persons. In calculation of age-adjusted death rates, there was no satisfactory method of selecting a threshold, or cut-off point, for number of deaths per population below which calculations would not be done. Therefore, in those cases where a few deaths occur in a very small population, the age-adjusted death rate may be extremely high. The user may avoid problems with interpretation of the data by glancing at population data to obtain a feeling for the distribution across race-sex-age groups.

There were isolated cases in which one or more deaths occurred in a county for a race-sex-age group for which there were no people. This problem may be due to the inherently imprecise nature of Census data when working with small numbers or to inaccuracies in the NCHS tapes. In these cases, the age-specific rate for that race-sex-age group was set to 0.

When all loading and verification of the mortality data were completed, all components in the mortality repeating group were converted back to KEY'ed components. The data base was then reorganized using the REORGANIZE ALL command to re-structure the data base tables and pointers to maximize efficiency of retrievals.

The 1970 age-specific population data were taken from raw Census tapes and aggregated into 15 age groups and 8 race-sex categories (4 total categories and 4 urban categories). It was assumed that since age-specific Census data would be accessed rather infrequently, those race-sex categories defined in the data base would be the minimum necessary for derivation of all categories. For example, total non-white can be derived by subtracting total white from the total population. User-defined functions are provided to derive those components not explicitly defined.

The Census data were aggregated, formatted for entry into the data base, and then placed in mass storage files. A PLI program was used to load the data by means of the PLI optimized loading. During loading, another component for non-white males was calculated and loaded to further simplify calculations of components not explicitly defined for use in in-out migration and mortality calculations. To further enhance the Census data repeating group, figures were totalled across the 15 age groups to create a 16th data set giving county totals for each component. These totals were loaded as data with AGE GROUPING code 0.

Geocoding problems with independent cities for Census data were handled as has been previously discussed. The Census county codes are shown in the cross-reference file of the POPATRISK Geocoding Manual except for the District of Columbia. There were no data for the District of Columbia on the Census tape, and therefore, the data had to be manually extracted from Census publication PC(V2)-10 General Population Characteristics and loaded separately.

Verification of the Census data has been an on-going process of spot cross-checking against various sources. The primary source used for verification was Census publication PC(V2) General Population Characteristics giving detailed population figures for each county. Loaded Census data were checked against PC(V2) and then cross-checked against the county entry level data and the Bureau of Census publication 1972 County and City Data Book (CCDB) to assure data loaded into each county were the correct data for that county.

There were found, in some cases, to be minor discrepancies between the total age-specific population as shown in the Census repeating group and the total population as shown in component C10. These discrepancies are explained in the CCDB to occur because figures shown in the main body of the book are only representative counts. Complete counts appear in Appendices A and B of the CCDB and correspond exactly with those loaded into the data base Census repeating group. In order to resolve the problem and to eliminate confusion over the figures, C10 TOTAL POPULATION figures were changed to show complete counts.

Conversion to SYSTEM 2000 Version 2.80

Loading of the data base was initiated using SYSTEM 2000 Version 2.65. The data base was partially loaded when the National Computer Center installed a new version of SYSTEM 2000 Version 2.80. Normally, new releases of the system are fairly transparent to data base development and require a minimal effort to adapt to changes. However, Version 2.80 contains major changes to the basic file structure, namely decreasing page size from 504 words to 448 words. This change, in effect, reduces input/output charges for accessing and updating the data base. A reload of the data base was necessary to gain optimal benefit from Version 2.80 features. Reloading also provided an opportunity to make desired modifications to the data base design to provide for more efficient retrievals and space in the data base for newly obtained in-out migration and mortality data.

Redesign and conversion to Version 2.80 involved unloading the existing data base and reloading the data into a newly designed data base created under Version 2.80. The conversion went smoothly, and the benefits of the redesign and conversion far outweighed the cost of the effort.

USE OF REMOVABLE DISK PACK AND ARCHIVING PROCEDURES

As data base development proceeded and the size of the data base files began to increase substantially, the cost of maintaining the files on mass storage became prohibitive. The use of magnetic tapes as the primary storage medium was also undesirable because of the high I/O costs involved in transferring large files between tape and mass storage.

In light of the significant increase in file sizes expected with the addition of in-out migration and mortality data to the data base, the most cost-effective solution to the data base storage problem was a private removable disk pack. Therefore the POPATRISK data base was transferred to removable disk pack to minimize monthly charges for disk space utilization and to provide for relatively easy access for frequent use.

There are a few drawbacks to use of the private disk pack which should be mentioned here. The first is the inconvenience of having to have the disk pack mounted before accessing the data base. Experience has shown, however, that this is only a minor inconvenience and usually involves only a 3-5 minute wait. The other drawback is that files on the disk pack are not protected from accidental loss by the Secure Processor. Archiving of files residing on a disk pack is left completely to the user.

Utilization of a dependable backup and restore system is vital to successful creation of a SYSTEM 2000 data base. SYSTEM 2000 flags a data base damaged whenever an update job, either demand or batch, terminates abnormally. Having been saved prior to the update job, the previous updated version of the data base may be restored and operations resumed. Since this error is generated by an update run 'MAX TIME' or any EXEC 8 system interrupt,it became a fairly common problem with POPATRISK and was resolved routinely.

A POPATRISK backup system was utilized during development of the data base to both insure against accidental loss and to provide sufficient archiving of data base updates. Similar to backup systems used by the National Air Data Branch of EPA, n generations of the data base are archived on tape, one generation per backup tape. The tapes are cycled each time the data base is saved so that only the last n archival versions of the data base are retained. An archival restore runstream was used in conjunction with the backup system to provide access to previous versions of the data base. Both backup and restore runstreams are fully documented in Appendix B.

A UNIVAC EXEC 8 system error encountered during data base development should be mentioned here because of its impact on the data base and the substantial cost in man hours and computer time required for recovery. National Computer Center personnel determined that a system error caused damage to one or more of the data base files, resulting in a shift in data within the data base. Since the data base was not flagged as damaged, the error went undetected throughout a significant portion of data base loading and archiving. Recovery from the damage required recovering appropriate data files and reloading all data loaded since the last correct version of the data base. This problem of unflagged damage to data is a data processor's nightmare and is mentioned here to emphasize the point that although safeguards are built into a data management system, they do not always work. Experience with POPATRISK has shown the importance of continuous monitoring of existing data in the data base as well as extensive verification of data presently being loaded.

The POPATRISK data base was loaded by both S2K Natural Language and Procedural Language COBOL Interface programming. During the initial stages of data base development,when moderate amounts of data were being loaded and data base files were small, Natural Language Queue Access loading performed a satisfactory function. As the volume of incoming data grew substantially and the data base files expanded, increases in computer I/O changes made Natural Language processing prohibitive. Considering the need for a more efficient method of loading large quantities of data and the necessity for interaction with data base data during processing, and loading, it was decided that the benefits of PLI programs would far out-weigh the costs of their development.

Procedural Language Interface optimized loading provided a much more cost-effective method of loading and allowed full access to existing data base data, a capability most useful in calculation of migration and age-adjusted mortality data. However, even with the increased efficiency of loading, by the time mortality data were being loaded, the data base files had expanded such that running times and I/O charges for mortality load jobs were growing exhorbitantly.

The major factor contributing to the increased I/O charges was the number of data sets to be loaded, all components of which were KEY'ed. A decision was made to make the components NON-KEY during loading, a new option available under SYSTEM 2000 Version 2.80. The action in fact decreased mortality data load times by at least a factor of 10. Changing the components back to KEY upon completion of loading was accomplished in 2 hours of computer time. The net results of the loading procedure were substantial savings in time and computer charges. The REORGANIZE ALL command was issued at various times during loading and upon completion of data base development to reorder all the data base tables whose entries were scattered as a result of incremental load and update operations.

APPENDIX A

RESOLUTION OF GEOCODING DISCREPANCIES FOR MASSACHUSETTS NEDS AND SAROAD DATA

Overview of Work Done on Massachusetts NEDS

Massachusetts NEDS point sources were assigned to appropriate Census counties by a computerized procedure involving geographic location by UTM coordinates. Area sources were apportioned by first aggregating the entire state's area emissions and then disaggregating into Census counties following the EPA publication, "Guide for Compiling a Comprehensive Emission Inventory." Massachusetts SAROAD monitoring stations were assigned by identifying the city name of all SAROAD monitoring sites and then locating the Census county in which the city was located.

Detailed Procedures Used in Treating NEDS Point Sources

To obtain Census county emissions totals for NEDS point sources in Massachusetts, the points were apportioned from NEDS counties into Census counties. SSI staff accessed the AEROS NADB*NEDS-USER file to retrieve the following data for approximately 2500 Massachusetts points.

STATE
COUNTY
AQCR
PLANT ID NUMBER
CITY
UTM ZONE
YEAR OF RECORD
ESTABLISHMENT NAME AND ADDRESS
POINT-ID
UTM-HORIZONTAL
UTM-VERTICAL
EMISSION ESTIMATES OF PARTICULATES
EMISSION ESTIMATES OF SO_2
EMISSION ESTIMATES OF NO_X
EMISSION ESTIMATES OF HC
EMISSION ESTIMATES OF CO

These data were supplied to an SSI consultant, Dr. Richard J. Kopec, who identified Census counties for each point source according to UTM coordinates of the point. For the several points in which UTM coordinates were missing in the NEDS-USER file, SSI staff searched for another point in the file with the same city code as the point with incomplete information. In most cases, the UTM coordinate of this new point was used to locate the original point in the correct Census county.

For example, the NEDS point

STATE	22
COUNTY	1291
AQCR	119
PLANT ID NUMBER	0510
CITY	1700
UTM ZONE	19
YEAR OF RECORD	72
ESTABLISHMENT NAME AND ADDRESS	POLAROID, 1 UPLAND RD,
POINT-ID	03
UTM-HORIZONTAL	
UTM-VERTICAL	
EMISSION ESTIMATES OF PARTICULATES	0000000
EMISSION ESTIMATES OF SO_2	0000000
EMISSION ESTIMATES OF NO_x	0000000
EMISSION ESTIMATES OF HC	0000008
EMISSION ESTIMATES OF CO	0000000

has no UTM coordinates (UTM-HORIZONTAL, UTM-VERTICAL). It has a city code of 1700. The following point, also with a city code of 1700, and with UTM coordinates present in the point record was located in the NEDS-USER FILE.

STATE	22
COUNTY	1291
AQCR	119
PLANT ID NUMBER	5732
CITY	1700
UTM ZONE	19
YEAR OF RECORD	72
ESTABLISHMENT NAME AND ADDRESS	RICHARD A KLEIN CO.349 LENOX,NORWOO
POINT-ID	01
UTM-HORIZONTAL	3181
UTM-VERTICAL	46723
EMISSION ESTIMATES OF PARTICULATES	0000000
EMISSION ESTIMATES OF SO_2	0000000
EMISSION ESTIMATES OF NO_x	0000000
EMISSION ESTIMATES OF HC	0000022
EMISSION ESTIMATES OF CO	0000000

The UTM coordinates of this source were substituted in the point record of the original source. These "dummy" UTM coordinates were used to place the original source in the correct Census county. Table A-1 lists the point sources with missing UTM coordinates and includes the "dummy" coordinate used to identify the point's Census county. Although in many cases identical multiple entries appear in the NEDS file due, e.g., to multiple stacks at a given site, only one point is listed below for each plant. (Points around the city of Boston were carefully checked when using this method of Census county identification because the city lies in more than one county boundary.)

In a few instances it was not possible to locate other points in the file with the same city code as the incomplete point source. For example, no points in the Lower Pioneer Valley Region (City Code 0860) contained UTM coordinates in the point records, so there was no "dummy" UTM coordinate available. Also, several point records contained UTM coordinates that misplaced the point source by locating it outside of the state.

In these cases the name and address of each point was traced to place the point in the correct Census county. Table A-2 contains record descriptions of point sources which were located in Census counties by this method. Again, only one set of values is listed for each plant.

After identifying Census counties for all the Massachusetts points, county totals were calculated for each of the five emission categories. Table A-3, "Total Emissions Summed by Type and by County," illustrates the results of this summation and the data that were loaded into the POPATRISK data base.

Detailed Results for NEDS Area Sources

Table A-4 gives the results of apportioning state total emissions among 14 Census counties.

Table A-1

NEDS POINTS APPORTIONED INTO CENSUS COUNTIES BY USING "DUMMY" UTM COORDINATES

STATE	COUNTY	AQCR	PLANT ID NUMBER	CITY	UTM ZONE	YEAR OF RECORD	NAME AND ADDRESS	POINT-ID	DUMMY UTM-HORIZONTAL	DUMMY UTM-VERTICAL	EMISSION ESTIMATES OF PARTICULATES	EMISSION ESTIMATES OF SO_2	EMISSION ESTIMATES OF NO_x	EMISSION ESTIMATES OF HC	EMISSION ESTIMATES OF CO
22	1291	119	0184	1880	19	70	GENERAL DYNAMICS,97 E HOWARD ST, QUINCY	06	3370	46785	0000000	0000000	0000000	0000763	0000000
22	1291	119	0510	1700	19	72	POLAROID, 1 UPLAND RD,	03	3181	46723	0000000	0000000	0000000	0000067	0000000
22	1291	119	5708	2340	19	72	J L THOMPSON RIVET, SAWYER, WALTHAM	02	3138	46923	0000000	0000000	0000000	0000007	0000000
22	1291	119	5905	0380	19	72	FLEXTEX, 224 CRESCENT AVE,	01	3315	46949	0000000	0000000	0000000	0000050	0000000
22	1291	119	5910	1780	19	72	S+S LEATHER, 150 MAIN ST,	01	3422	47093	0000000	0000000	0000000	0000179	0000000
22	1291	119	6744	0360	19	72	AMERICAN LAUND, 57 MT AUBURN,CAMBRIDGE	01	3267	46932	0000000	0000000	0000000	0000008	0000000
22	1291	119	7230	2380	19	72	OPELL CO, 60 ACTION ST,	01	3226	46928	0000000	0000000	0000000	0000259	0000000
22	1291	119	7253	0660	19	72	GM ASSEMBLY DIV GMC, 63WESTRN,FRAMGHAM	02	3010	46813	0000000	0000000	0000000	0002132	0000000
22	1291	119	7292	0480	19	72	SYLVANIA, 100 ENDICOTT ST, DANVERS	01	3383	47137	0000000	0000000	0000000	0000568	0000000
22	1291	119	7513	0240	19	72	AMERICAN ARCHT, 95 BORDER ST,	01	3230	46912	0000000	0000000	0000000	0000009	0000000
22	1291	119	7524	1220	19	72	CONTAINER CORP AM, 200 BOSTON,	01	3257	46973	0000000	0000000	0000000	0000029	0000000
22	1291	119	7529	1480	19	72	DAMON, 300 SECOND AVE, NEEDHAM	01	3156	46844	0000000	0000000	0000000	0000026	0000000
22	1291	119	7558	0350	19	72	ITEK, 154 MIDLSX TNPK, BURLINGTON	01	3178	47050	0000000	0000000	0000000	0000005	0000000
22	1291	119	7700	0240	19	72	MOBIL OIL CO, CHELSEA ST, E BOSTON	01	3230	46912	0000000	0000000	0000000	0001359	0000000
22	1291	119	7701	0240	19	72	WHITE FUEL, S BOSTON TRMNAL, S BOSTON	01	3230	46912	0000000	0000000	0000000	0000202	0000000
22	1291	119	7702	0380	19	72	GULF OIL CO, CHELSEA TRMNAL, CHELSEA	01	3315	46949	0000000	0000000	0000000	0000338	0000000
22	1291	119	7704	0560	19	72	HUMBLE OIL, 151 BOW ST, EVERETT	01	3299	46949	0000000	0000000	0000000	0000906	0000000
22	1291	119	7705	0260	19	72	CITIES SERVCE OIL,345 QUINCY AV,E BRNTR	01	3327	46736	0000000	0000000	0000000	0001850	0000000
22	1291	119	7706	1180	19	72	MOBIL OIL, 279 MAPLE ST, MARLBORO	01	2896	46931	0000000	0000000	0000000	0000018	0000000
22	1291	119	7707	1940	19	72	ATLANTIC RCHFLD,140 LEE BRBNK HWY,REVER	01	3360	46969	0000000	0000000	0000000	0000492	0000000
22	1291	119	7708	1940	19	72	NORTHEAST PETROL, LE BRBNK HWY, REVERE	01	3360	46969	0000000	0000000	0000000	0001240	0000000
22	1291	119	7709	1940	19	72	SUN OIL CO, LE BRBNK HWY, REVERE	01	3360	46969	0000000	0000000	0000000	0000422	0000000
22	1291	119	7710	1940	19	72	GIBBS OIL CO, 40 LEE BRBNK HWY, REVERE	01	3360	46969	0000000	0000000	0000000	0003894	0000000
22	1291	119	7712	2340	19	72	HUMBLE OIL CO, 50 JONES RD WALTHAM	01	3138	46923	0000000	0000000	0000000	0000244	0000000
22	1798	042	0208	0532	18	74	MILTON BRADLEY, 443 SHAKER RD,	01	7049	46585	0000000	0000000	0000000	0000049	0000000
22	2121	120	0142	1230	19	73	USM CORP, 163 MAIN, MEDWAY	01	3030	46680	0000000	0000000	0000000	0000004	0000000
22	2121	120	0185	2240	19	73	HASKON INC, 336 WEIR, TAUNTON	01	3251	42418	0000000	0000000	0000000	0000007	0000000
22	2121	120	0186	0507	19	73	ICI UNITED STATES, 333 MAIN, DIGHTON	01	3245	46355	0000000	0000000	0000000	0000006	0000000
22	2121	120	0187	0120	19	73	ROBBINS CO, ONEIL BLVD, ATTLEBORO	01	3109	46489	0000000	0000000	0000000	0000012	0000000
22	2121	120	0209	0120	19	73	ENGLEHARD IND, DUNHAM ST, ATLBORO	01	3109	46489	0000000	0000000	0000000	0000024	0000000
22	2121	120	0210	1810	19	73	ENGLEHARD IND, RT 152, PLAINVLE	01	3066	46522	0000000	0000000	0000000	0000020	0000000

Table A-2

NEDS POINTS APPORTIONED INTO CENSUS COUNTIES BY TRACING THEIR NAME AND ADDRESS

STATE	COUNTY	AQCR	PLANT ID NUMBER	CITY	UTM ZONE	YEAR OF RECORD	NAME AND ADDRESS	POINT-ID	DUMMY UTM-HORIZONTAL	DUMMY UTM-VERTICAL	EMISSION ESTIMATES OF PARTICULATES	EMISSION ESTIMATES OF SO_2	EMISSION ESTIMATES OF NO_x	EMISSION ESTIMATES OF HC	EMISSION ESTIMATES OF CO
							* POINTS WITH INACCURATE UTM COORDINATES *								
22	1291	119	0197	2100	19	70	EASTERN BAKING,30 INNER BELT RD,SOMRVILL	01	3533	46936	0000002	0000027	0000005	0000000	0000000
22	1291	119	5459	0240	19	72	STEDFAST RUBBER CO,34 TOPLAIN,BOSTON	01	2694	68121	0000000	0000000	0000000	0000034	0000000
22	1291	119	6694	1540	19	72	HOLLY CLNRS,724 BEACON ST,NEWTN CNTR	01	5193	40885	0000000	0000000	0000000	0000009	0000000
22	1291	119	6746	2380	19	72	G A CLEANERS, 475 MAIN ST, WATERTOWN	01	3194	69300	0000000	0000000	0000000	0000007	0000000
22	1291	119	6836	0823	19	72	HALLIDAY LITHO CORP, CIRCUIT, HANOVER	01	3490	44640	0000000	0000000	0000000	0000013	0000000
22	1291	119	7539	1900	19	72	ELLIOT BUSINS MACH, IND PK RANDOLPH	01	2304	46723	0000000	0000000	0000000	0000005	0000000
22	2121	120	0026	0744	19	70	GOSNOLD DUMP, GOSNOLD	01	3380	45870	0000001	0000000	0000000	0000001	0000002
22	1798	042	0205	2160	18	74	MONSANTO, 190 GROCHMAL	01	7050	44703	0000000	0000000	0000000	0000685	0000000
22	1798	042	0050	1384	18	70	MONROE DUMP, MONROE	01	6701	46321	0000004	0000000	0000001	0000008	0000022
							* POINTS WITH NO UTM COORDINATES *								
22	1798	042	0194	0860	18	74	LUDLOW CORP, 111 MOSHER ST, 01040	01			0000000	0000000	0000000	0000033	0000000
22	2121	120	0143	1340	19	73	BETTINGER CO, SUMMER, MILFORD	01			0000000	0000000	0000000	0000015	0000000
22	2121	120	0144	1340	19	73	KNOTT MFG, FRONT, MILFORD	01			0000000	0000000	0000000	0000006	0000000
22	2121	120	0172	2680	19	73	WRENTHAM STEEL,KENDRICKE,WRENTHAM	01			0000000	0000000	0000000	0000005	0000000
22	2121	120	0177	1340	19	73	MILFORD SHOE, 62 NORTH BOW, MILFORD	01			0000000	0000000	0000000	0000004	0000000
22	2121	120	0204	1340	19	73	CNTRL CLNRS, 145 BOW, MILFORD	01			0000000	0000000	0000000	0000005	0000000

Table A-3

TOTAL EMISSIONS SUMMED BY TYPE AND BY COUNTY

```
        COUNTY=1                       COUNTY=11                      COUNTY=21
       Barnstable                      Franklin                        Norfolk

VARIABLE         SUM          VARIABLE         SUM          VARIABLE         SUM

PART      1383.0000000        PART       500.00000000       PART      1080.0000000
SO2      31682.0000000        SO2       1297.00000000       SO2       3905.0000000
NOX       8699.0000000        NOX        328.00000000       NOX        939.0000000
HC         590.0000000        HC         731.00000000       HC       16457.0000000
CO        1202.0000000        CO        1409.00000000       CO         750.0000000

-------COUNTY=3--------       -------COUNTY=13--------      ------COUNTY=23--------
       Berkshire                       Hampden                        Plymouth

PART      1363.0000000        PART      27003.0000000       PART       173.00000000
SO2      10727.0000000        SO2       55770.0000000       SO2         11.00000000
NOX       2017.0000000        NOX       22517.0000000       NOX         64.00000000
HC        1178.0000000        HC        18852.0000000       HC         928.00000000
CO        2192.0000000        CO         3433.0000000       CO         922.00000000

-------COUNTY=5--------       -------COUNTY=15--------      ------COUNTY=25--------
        Bristol                        Hampshire                      Suffolk

PART      6726.000000         PART       9260.0000000       PART      3325.0000000
SO2     109956.000000         SO2       22639.0000000       SO2       7324.0000000
NOX      36436.000000         NOX        4915.0000000       NOX       1658.0000000
HC       23459.000000         HC         8407.0000000       HC       38277.0000000
CO        2858.000000         CO         1633.0000000       CO         122.0000000

-------COUNTY=7--------       -------COUNTY=17--------      ------COUNTY=27--------
         Dukes                         Middlesex                      Worcester

PART        44.00000000       PART       5176.0000000       PART     14437.0000000
SO2          2.00000000       SO2       12562.0000000       SO2      19081.0000000
NOX         15.00000000       NOX        2860.0000000       NOX       9228.0000000
HC          30.00000000       HC        28724.0000000       HC        6071.0000000
CO         227.00000000       CO         3248.0000000       CO       17810.0000000

-------COUNTY=9--------       -------COUNTY=19--------
         Essex                         Nantucket

PART      2344.0000000        PART        86.00000000
SO2      15274.0000000        SO2          5.00000000
NOX       3859.0000000        NOX         32.00000000
HC        7246.0000000        HC         230.00000000
CO        1128.0000000        CO         459.00000000
```

Table A-4

MASSACHUSETTS AREA EMISSIONS TOTALS BY POPATRISK COUNTIES

(All State 22)

POPATRISK COUNTY	TSP	SO_2	NO_X	CO
0001	1,136.5	1,827.2	6,558.1	56,345.0
0003	1,806.6	3,747.5	9,068.3	67,896.0
0005	4,711.1	11,095.3	21,000.7	138,511.3
0007	70.2	102.4	426.0	3,785.2
0009	6,741.9	14,958.0	31,680.4	223,638.6
0011	580.3	1,280.0	2,689.4	18,994.2
0013	4,989.0	10,882.6	23,954.8	172,158.6
0015	1,104.6	2,574.9	5,046.8	34,182.5
0017	13,872.5	30,154.6	66,601.1	486,865.6
0019	47.8	93.6	242.1	1,931.7
0021	3,718.8	12,400.9	27,461.9	198,855.1
0023	3,127.9	7,097.2	14,495.6	100,580.0
0025	6,520.5	14,569.4	29,334.6	204,053.0
0027	5,582.6	15,299.7	29,877.5	201,280.0

APPENDIX B

DOCUMENTATION OF DATA TAPES AND RUNSTREAMS

BUREAU OF THE CENSUS IN-OUT MIGRATION TAPE FILE DESCRIPTION

1. File Title: Migration by counties, 1970
2. Technical Characteristics
   a. Tape type and density

      9 track, 1600 BPI, EBCDIC, odd parity, standard IBM OS labels.

   b. Record/field size. Fixed length

      1464 character records, 7320 character blocks. Data records contain 120 characters for geography and 168 8-character data cells.
3. File Size: 3192 records
4. File Sequence: FIPS state by FIPS county.
5. File contains one record for each state and one record for each county. State records contain a "000" for county code and a record identifier of "A" County records contain a valid county code and a blank record identifier. There is a special record for New York City as a whole in addition to the individual counties in New York City. This record is identified as state 36, county 000 and record identifier "B."

   There is also an additional out-migrant record in New York coded 36995 which contains the tabulation of only records with a response of "New York City--no borough given" for 1965 and a residence code in 1970 outside of New York City. This out-migrant record, however, should not be used with much confidence.

RECORD LAYOUT

Matrix is:

Migration (2), Race (3), Sex (2), Age (7), Allocation (2). (168 cells).

Geography	Characters
State	1 - 2
County	3 - 5
Identifier	6
Unused	7 - 119
Dollar Sign	120

Data	121 - 1464
Immigration	121 - 792
Total All Races	121 - 344
Male	121 - 232
5-14 years	
Allocated	121 - 128
Not Allocated	129 - 136
15-19 years	
Allocated	137 - 144
Not Allocated	145 - 152
20-24 years	
Allocated	153 - 160
Not Allocated	161 - 168
25-29 years	
Allocated	169 - 176
Not Allocated	177 - 184
30-34 years	
Allocated	185 - 192
Not Allocated	193 - 200
45-64 years	
Allocated	201 - 208
Not Allocated	209 - 216
65 years and over	
Allocated	217 - 224
Not Allocated	225 - 232
Female	233 - 344
5-14 years	
Allocated	233 - 240
Not Allocated	241 - 248
15-19 years	
White	345 - 568
Male	345 - 456
⋮	
Female	457 - 568
⋮	
Black	
Male	569 - 680
⋮	
Female	681 - 792
⋮	

Out Migration	
Total all races	
Male	793 - 904
⋮	
Female	905 - 1016
White	
Male	1017 - 1128
⋮	
Female	1129 - 1240
⋮	
Black	
Male	1241 - 1352
⋮	
Female	1353 - 1464
⋮	

In-migrants in this file also are persons who reported a foreign country as their residence in 1965; however, the out-migrant category excludes persons who had reported their residence in that county in 1965 but were living overseas in 1970. For counties that do not have any persons reporting a foreign county as their residence in 1965, the count of unallocated in-migrants will equal the number shown in Table 119 of the PC(1)-C report.

6. Considerations Regarding File Content

Allocation Method:

The data presented in this file are based on question 19 of the 1970 Census 15 percent household sample questionnaire. It requested those persons who reported living in a different house in 1965 to report the state, county and city, town or village (if applicable) for their residence in 1965. Persons who indicated that they lived in a different house but did not report its address were classified as "moved, residence not reported."

An allocation procedure was developed for such nonresponses which assigned a 1965 state and county of residence based on the respondents age, sex, race, college status and military status. Such allocations are by design valid only to the county level of geography.

To allow comparability with published reports, both allocated and unallocated migrant counts are shown for each county or Census county division in the United States.

Suppression

Data on in-migrants were suppressed for those counties that had fewer than 34 weighted cases. Specifically, these counties are:

1) Angoon County, Alaska (St. 02, County 030)
2) Hinsdale County, Colorado (St. 08, County 055)
3) Yellowstone National Park County, Montana (St. 30, County 113)

Data fields for these counties will contain zeroes.

Comparison with Published Data

In theory, the unallocated counts of migrants into any county developed from this file should equal the number of persons living in a different county plus those abroad in 1965. (See Table 119, 125, or 145 of the 1970 Census State reports.)

However, there will be in some counties fewer unallocated "different county in-migrants" than are shown in the published reports. This stems from the allocation procedure which altered the records for persons who were in the military in either 1965 or 1970 and who reported "at sea in 1965" by substituting a county of residence in 1965. However, if the individual was not in the military either in 1965 or 1970, his "at sea" response was not changed.

Some counties may also have more unallocated "different county in-migrants" than are shown in the published reports. This stems from the different interpretations of a complete report for persons reporting a foreign country as their residence in 1965.

The number of unallocated out-migrants shown on this file for a given county may be compared to Table 2 of the PC(2)-2E Migration Between State Economic Areas report provided that the given county is also a state economic area.

Migration Between New York City Boroughs

A special code was assigned to those persons living in New York City in 1970 whose reponse for 1965 residence merely indicated different house-New York City. Persons assigned this code were not included in either the in- or out-migrant tallies for individual boroughs.

Because of this unique situation, a special record is provided for New York City as a whole. (This record is identified as State 36, County 000 and Record Identifier B.) For this summary record, in-migrants exclude those persons who reported their residence in 1965 to either a New York City borough and those persons cited to "place of residence-New York City-no borough given."

In- and out-migrants for specific boroughs include only those persons with either allocated or complete responses to residence five years ago. Persons giving "New York City-no borough given" were not included as in- or out-migrants from boroughs.

MORTALITY WORKTAPE FILE DESCRIPTION-AGGREGATION OF 1969-71 MORTALITY DATA INTO 50 ICDA CATEGORIES

1. File Title: Mortality Counts for 50 ICDA's

2. Technical Characterisitics

   a. Tape type and density

      9 track, 1600 BPI, standard UNIVAC labeled tape

   b. Record/field size. Fixed length

      24005 character records, 24005 character blocks. Data records contain 5 characters for state and county and 3000 8 character data cells.

3. File size: 3067 records

4. File sequence. File contains one record for each NCHS county, one record provided for New York City area as a whole with no borough breakdown. Records for Virginia independent cities have been combined with their original county records. Each record contains combined mortality counts for 1969-1971 broken out into 50 categories (48 HERL categories + 2 approved neoplasm clusters), 15 age groups, and 4 race-sex categories.

RECORD LAYOUT

Matrix is:

ICD # (50), Age (15), Sex Race (4), (3000 cells).

Geography	Characters
State	1 - 2
County	3 - 5
Data	6 - 24005
ICDA - 1	
0-4 Age Group	
White Male	6 - 13
White Female	14 - 21
Non-white Male	22 - 29
Non-white Female	30 - 37

Geography	Characters
5-9 Age Group	
White Male	38 - 45
White Female	46 - 53
Non-white Male	54 - 61
Non-white Female	62 - 69
10-14 Age Group	70 - 101
White Male	
White Female	
Non-white Male	
Non-white Female	
15-19 Age Group	102 - 133
White Male	
White Female	
Non-white Male	
Non-white Female	
20-24 Age Group	134 - 165
White Male	
White Female	
Non-white Male	
Non-white Female	
25-29 Age Group	166 - 197
White Male	
White Female	
Non-white Male	
Non-white Female	
30-34 Age Group	198 - 229
White Male	
White Female	
Non-white Male	
Non-white Female	
35-39 Age Group	230 - 261
White Male	
White Female	
Non-white Male	
Non-white Female	
40-44 Age Group	262 - 293
White Male	
White Female	
Non-white Male	
Non-white Female	

Geography	Characters
45-49 Age Group	247 - 325
White Male	
White Female	
Non-white Male	
Non-white Female	
50-54 Age Group	326 - 357
White Male	
White Female	
Non-white Male	
Non-white Female	
55-59 Age Group	358 - 389
White Male	
White Female	
Non-white Male	
Non-white Female	
60-64 Age Group	390 - 421
White Male	
White Female	
Non-white Male	
Non-white Female	
65-69 Age Group	422 - 453
White Male	
White Female	
Non-white Male	
Non-white Female	
70 > Age Group	454 - 485
White Male	
White Female	
Non-white Male	
Non-white Female	
ICDA - 2 through ICDA - 50	
0-4 Age Group through 70 > Age Group	486 - 24005
White Male	
White Female	
Non-white Male	
Non-white Female	

POPATRISK DATA BASE BACKUP SYSTEM

A data base backup system was used during development of the POPATRISK data base to provide adequate archiving of updates during loading. Backup procedures were used to protect against computer system failures as well as data processing errors. The runstream documented below causes cycling of a specified number of backup tapes so that the current version of the data base is copied to the tape containing the oldest version.

```
@RUN
@SYM PRINT$,,'SITEID'
@ASG,T T.,F40,DSD087
@REMARK: . GET CYCLED BACKUP TAPE AND ASSIGN
@ED,R APPAR*POPTAPES.
OFF DSPLIT
SPLIT TAPE N N
@ADD,P TAPE
@REMARK: . COPY EACH DATA BASE FILE TO TAPE
@COPY,GM 1POPATRISK--.,POPSAVE.
@COPY,GM 2POPATRISK--.,POPSAVE.
@COPY,GM 3POPATRISK--.,POPSAVE.
@COPY,GM 4POPATRISK--.,POPSAVE.
@COPY,GM 5POPATRISK--.,POPSAVE.
@COPY,GM 6POPATRISK--.,POPSAVE.
@COPY,GM 7POPATRISK--.,POPSAVE.
@PRT,F 1POPATRISK--.
@PRT,F 2POPATRISK--.
@PRT,F 3POPATRISK--.
@PRT,F 4POPATRISK--.
@PRT,F 5POPATRISK--.
@PRT,F 6POPATRISK--.
@PRT,F 7POPATRISK--.
@ED,U APPAR*POPTAPES.
MOVE N
LNPRINT!
@FREE APPAR*POPSAVE
@FIN
```

```
@RUN
```

Data base backup requires just over five minutes SUP time. Therefore, the run card should specify 10 minutes run time.

```
@SYM PRINT$,,'SITEID'
```

This command is used to route output to user's remote terminal. Elimination of the command causes printing at the RTP on-site printer.

```
@ASG,T T.,F40,DSD087
```

A tempory file is assigned on the disk pack to assure pack is mounted prior to accessing files.

```
@ED,R APPAR*POPTAPES.
OFF DSPLIT
SPLIT TAPE N N
```

The UNIVAC editor is invoked to access a file containing a list of backup tape assignment commands. The SPLIT command copies the Nth tape assignment command to the temporary element TAPE.

```
@ADD,P TAPE
```

Temporary element TAPE is added into the runstream causing assignment of the selected tape.

```
@COPY,GM 1POPATRISK--.,POPSAVE.
@COPY,GM 2POPATRISK--.,POPSAVE.
@COPY,GM 3POPATRISK--.,POPSAVE.
@COPY,GM 4POPATRISK--.,POPSAVE.
@COPY,GM 5POPATRISK--.,POPSAVE.
@COPY,GM 6POPATRISK--.,POPSAVE.
@COPY,GM 7POPATRISK--.,POPSAVE.
```

Each of the seven SYSTEM 2000 data base files is copied to tape in COPY,G format.

```
@PRT,F 1POPATRISK--.
@PRT,F 2POPATRISK--.
@PRT,F 3POPATRISK--.
@PRT,F 4POPATRISK--.
@PRT,F 5POPATRISK--.
@PRT,F 6POPATRISK--.
@PRT,F 7POPATRISK--.
```

The Master File Directory entry for each file currently on the disk pack is printed

```
@ED,U APPAR*POPTAPES.
MOVE N
LNPRINT!
```

The tape file is edited to move the tape command just used to the top of the tape list, thus cycling the backup tapes.

```
@FREE APPAR*POPSAVE
@FIN
```

The backup tape is released and the run terminated.

The runstream documented below is used in conjunction with the backup runstream discussed above, and is used to copy a previous version of the data base from tape to disk pack.

```
@RUN
@SYM PRINT$,,'SITEID'
@ASG,T T.,F40,DSD087
@ASG,T POPSAVE.,16N,TAPE #
@DELETE,C 1POPATRISK--
@ASG,C 1POPATRISK--,///45000,DSD087
@COPY,G POPSAVE.,1POPATRISK--.
@DELETE,C 2POPATRISK--
@ASG,C 2POPATRISK--,///45000,DSD087
@COPY,G POPSAVE.,2POPATRISK--.
@DELETE,C 3POPATRISK--
@ASG,C 3POPATRISK--,///45000,DSD087
@COPY,G POPSAVE.,3POPATRISK--.
@DELETE,C 4POPATRISK--
@ASG,C 4POPATRISK--,///45000,DSD087
@COPY,G POPSAVE.,4POPATRISK--.
@DELETE,C 5POPATRISK--
@ASG,C 5POPATRISK--,///45000,DSD087
@COPY,G POPSAVE.,5POPATRISK--.
@DELETE,C 6POPATRISK--
@ASG,C 6POPATRISK--,///45000,DSD087
@COPY,G POPSAVE.,6POPATRISK--.
@DELETE,C 7POPATRISK--
@ASG,C 7POPATRISK--,F40/0/POS/999,DSD087
@COPY,G POPSAVE.,7POPATRISK--.
```

```
@FREE 1POPATRISK--
@FREE 2POPATRISK--
@FREE 3POPATRISK--
@FREE 4POPATRISK--
@FREE 5POPATRISK--
@FREE 6POPATRISK--
@FREE 7POPATRISK--
@PRT,F 1POPATRISK--.
@PRT,F 2POPATRISK--.
@PRT,F 3POPATRISK--.
@PRT,F 4POPATRISK--.
@PRT,F 5POPATRISK--.
@PRT,F 6POPATRISK--.
@PRT,F 7POPATRISK--.
@FREE POPSAVE.
@FIN
```

```
@RUN
```

Run card should contain 10 minutes run time.

```
@SYM PRINT$,,'SITEID'
@ASG,T T.,F40,DSD087
```

Same as in backup runstream.

```
@ASG,T POPSAVE.,16N,TAPE #
```

Assign tape containing version of data base to be restored.

```
@DELETE,C 1POPATRISK--
@ASG,C 1POPATRISK--,///45000,DSD087
@COPY,G POPSAVE.,1POPATRISK--.
@DELETE,C 2POPATRISK--
@ASG,C 2POPATRISK--,///45000,DSD087
@COPY,G POPSAVE.,2POPATRISK--.
@DELETE,C 3POPATRISK--
@ASG,C 3POPATRISK--,///45000,DSD087
@COPY,G POPSAVE.,3POPATRISK--.
@DELETE,C 4POPATRISK--
@ASG,C 4POPATRISK--,///45000,DSD087
@COPY,G POPSAVE.,4POPATRISK--.
@DELETE,C 5POPATRISK--
@ASG,C 5POPATRISK--,///45000,DSD087
@COPY,G POPSAVE.,5POPATRISK--.
@DELETE,C 6POPATRISK--
@ASG,C 6POPATRISK--,///45000,DSD087
@COPY,G POPSAVE.,6POPATRISK--.
@DELETE,C 7POPATRISK--
@ASG,C 7POPATRISK--,F40/0/POS/999,DSD087
@COPY,G POPSAVE.,7POPATRISK--.
```

Each of the seven disk pack files are deleted, reassigned on the disk pack, and the files copied from tape to disk pack.

```
@FREE 1POPATRISK--
@FREE 2POPATRISK--
@FREE 3POPATRISK--
@FREE 4POPATRISK--
@FREE 5POPATRISK--
@FREE 6POPATRISK--
@FREE 7POPATRISK--
```

Free each disk pack file so the Master File Directory can be updated with current file information.

```
@PRT,F 1POPATRISK--.
@PRT,F 2POPATRISK--.
@PRT,F 3POPATRISK--.
@PRT,F 4POPATRISK--.
@PRT,F 5POPATRISK--.
@PRT,F 6POPATRISK--.
@PRT,F 7POPATRISK--.
```

Print out new Master File Directory information on each data base file.

```
@FREE POPSAVE.
@FIN
```

Release backup tape and terminate run.

APPENDIX C

OTHER STUDIES

SURVEY OF 1973 MONITORING ACTIVITY

This section documents an initial survey made of monitoring stations active in 1973 by county in the contiguous United States for seven major air pollutants: TSP, SO_2, NO_2, CO, Ozone, Suspended Nitrates and Suspended Sulfates. Tabulations of levels of monitoring activity were made for all counties in the 48 contiguous states (Alaska and Hawaii were not considered). In addition, seven maps were produced, one for each pollutant considered, on which the number of monitoring stations is indicated by a corresponding number written in the county of interest of U.S. Department of Commerce Maps of the United States, stock number 0301-1895.

This survey was undertaken in order to:

1) Determine the geographical distribution of air monitoring activity for the seven pollutants of interest in an easily comprehended pictorial form.

2) Determine the percentages of U.S. population on a state and national basis residing in counties for which the air quality is monitored by numbers of stations falling into the following categories: at least 1, at least 2, at least 3, at least 5, at least 10.

3) Provide a sound basis for judgment on the extent to which methods for augmenting the existing air quality data should be investigated.

Methodology

Sources of Data--

The primary source of data relating to air quality monitoring sites was the publication, Directory of Air Quality Monitoring Sites Active in 1973 (Directory), EPA-450/2-75-006. 1973 was chosen as the year investigated because of the ready availability of this siting information. The Directory lists siting information, pollutant monitored, method used, and other data on a state but not strictly a county basis. In order to obtain an accurate count of monitoring sites by county it was necessary to keep a running tabulation while scanning by hand through the states listed. A computer program to perform this task was considered but rejected as being much more costly to develop and implement.

Population data for the year 1972 were considered sufficiently accurate and were obtained from the 1972 County and City Data Book. A listing (SAROAD file name NADB-PARMFL) of methods, by pollutant, in use for monitoring purposes was obtained from Mr. A. A. Slaymaker, Chief, Data Processing Section, National Air Data Branch. Another publication found useful in identifying quickly the location and names of counties was Office of Air Program's Federal Air Quality Control Programs, #AP-102.

Tabulation of Monitoring Methods Included in and Excluded from the Survey--

A number of stations listed in the Directory used methods which were deemed unacceptable or unsatisfactory for the purposes of this survey. For example, methods yielding results not directly relatable to air quality standards were deemed unsatisfactory. Although a count of all methods in use was kept, the results reported are only for those methods considered to yield reliable air quality data. In Table C-1 we tabulate, for each pollutant surveyed, a list of methods included in the count and alongside (if applicable) a list of methods excluded. In no case would inclusion of

TABLE C-1. LISTING OF MONITORING METHODS INCLUDED IN AND EXCLUDED FROM SURVEY

Pollutant	Methods Included	Methods Excluded
SO_2	flame photometric West-Gaeke colorimetric conductimetric pulsed-flourescent pararosanaline-sulfuric acid total sulfur flame photometric instrumental coulometric	polarographic Hydrogen Peroxide NaOH titration Sequential-conductimetric Hydrogen Peroxide
NO_2	Colorimetric-Lyshkow (mod.) coulometric chemiluminescence Sodium Arsenite - frit NASN Sodium Arnsenite - orifice TEA method - frit TGS method - frit TGS method - orifice Griess-Satzman	polarographic TEA Method-orifice
Sulfates	Colorimetric turbidimetric methyl thymol blue (ASTM) barium sulfate (ASTM)	
Nitrates	2-4 Xylenol reduction-diazo coupling specific ion electrode phenol-disulfonic acid ultraviolet-spectrophotometric	
Ozone	chemiluminescence ultraviolet DASIBI CORP total oxidant - 0.2 ($NO + NO_2$) total oxidant colorimetric neutral KI	
TSP	Hi-Vol Membrane Sampler millipore filter cassette	sticky paper soil index tape sampler bucket gravimetric nephelometer smoke shade cyclone
CO	nondispersive infra-red gas chromatographic dual isotope flourescence	detection tube catalytic comb-thermal detector

unacceptable and/or unsatisfactory methods in the count materially alter the results of the survey. We are pleased to acknowledge the assistance of Mr. L. J. Purdue of the Environmental Monitoring and Support Lab, Research Center, RTP, in assisting in the technical judgment of reliability and applicability of monitoring methods.

The Case of Massachusetts--

Massachusetts presented some special problems in performing this survey. As of the date of this report, no EPA guidelines exist for a clear cut separation of Massachusetts geography into counties. Many areas are specified as townships. The 1970 Census, however, divides Massachusetts into 14 counties and it was decided to use this division as the basis for deriving siting information. The Directory specified the AQCR within which monitoring stations were located. By determining the exact location of each monitoring station within an AQCR we were able to determine the corresponding Census county. Census data then provided the necessary population figures.

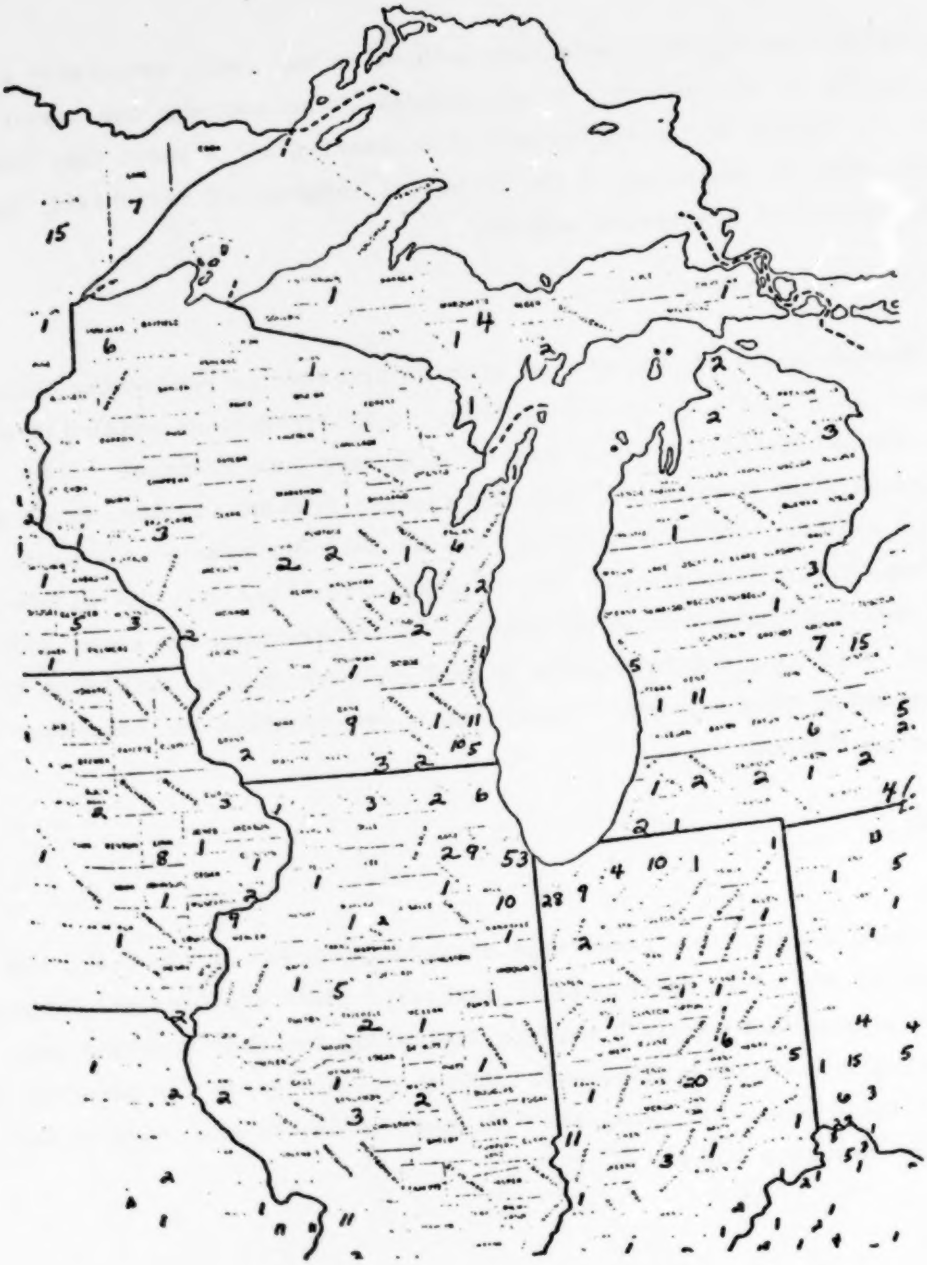
Results

Maps--

Figure C-1 presents seven maps of selected areas in the U.S., one for each of the seven pollutants of interest. Each map presents numbers written within each county's borders indicating how many stations are monitoring a particular pollutant. No numbers are shown for counties with no monitoring stations. Complete U.S. maps by county are being provided EPA under separate cover.

Tables--

Table C-2 shows tabulation tables, one for each of the seven air pollutants of interest, in which the results of this survey are summarized on a state basis. For each of the 48 states studied, the entries show the percentage of state population residing in counties (within the state) having at least 1, at least 2, at least 3, at least 5, and at least 10 stations monitoring that particular pollutant. Finally, corresponding total numbers for the 48 states are given.



Reproduced from best available copy.

Figure C-1. 1973 TSP Monitoring Station Sample Distribution

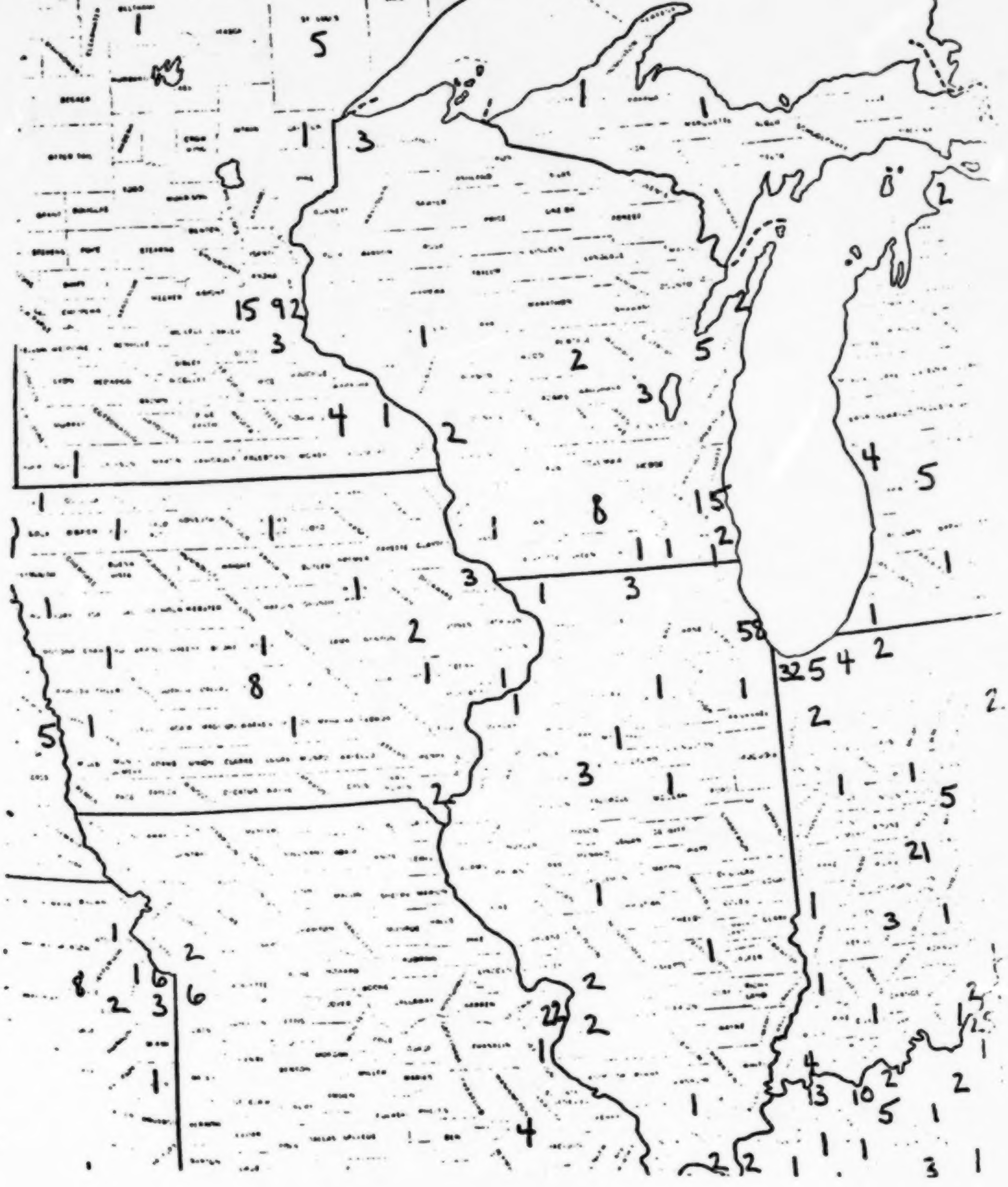

1
1
5
1
3
1
1
15 9 2
3
1
2
5
3
4
1
2
4
5
1
15
8
1
1 1
2
1
1
1
3
1
1
3
1
2
58
1
2
1
8
1
325 4
2
2
1
5
1
1
2
1
1
3
1
2
1
1
5
21
1
3
1
1
1
2
8
1 6
6
2 3
2
22
2
1
1
1
2
1
1
1
2
4
13
10 2
5
2
1
4
2 2
1
1
1
1
3
1

Figure C-1. 1973 SO_2 Monitoring Station Sample Distribution

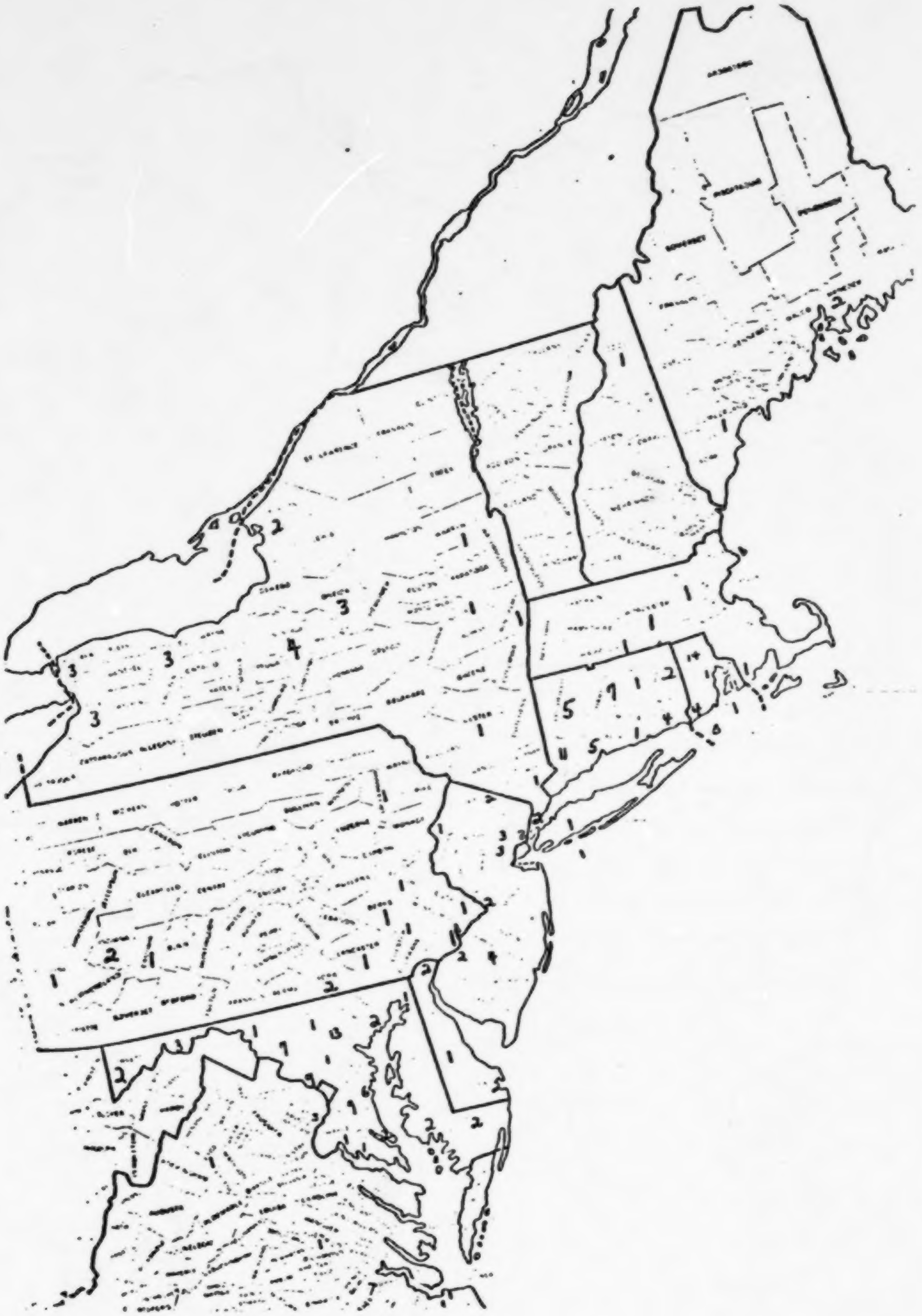


Figure C-1. 1973 NO_2 Monitoring Station Sample Distribution

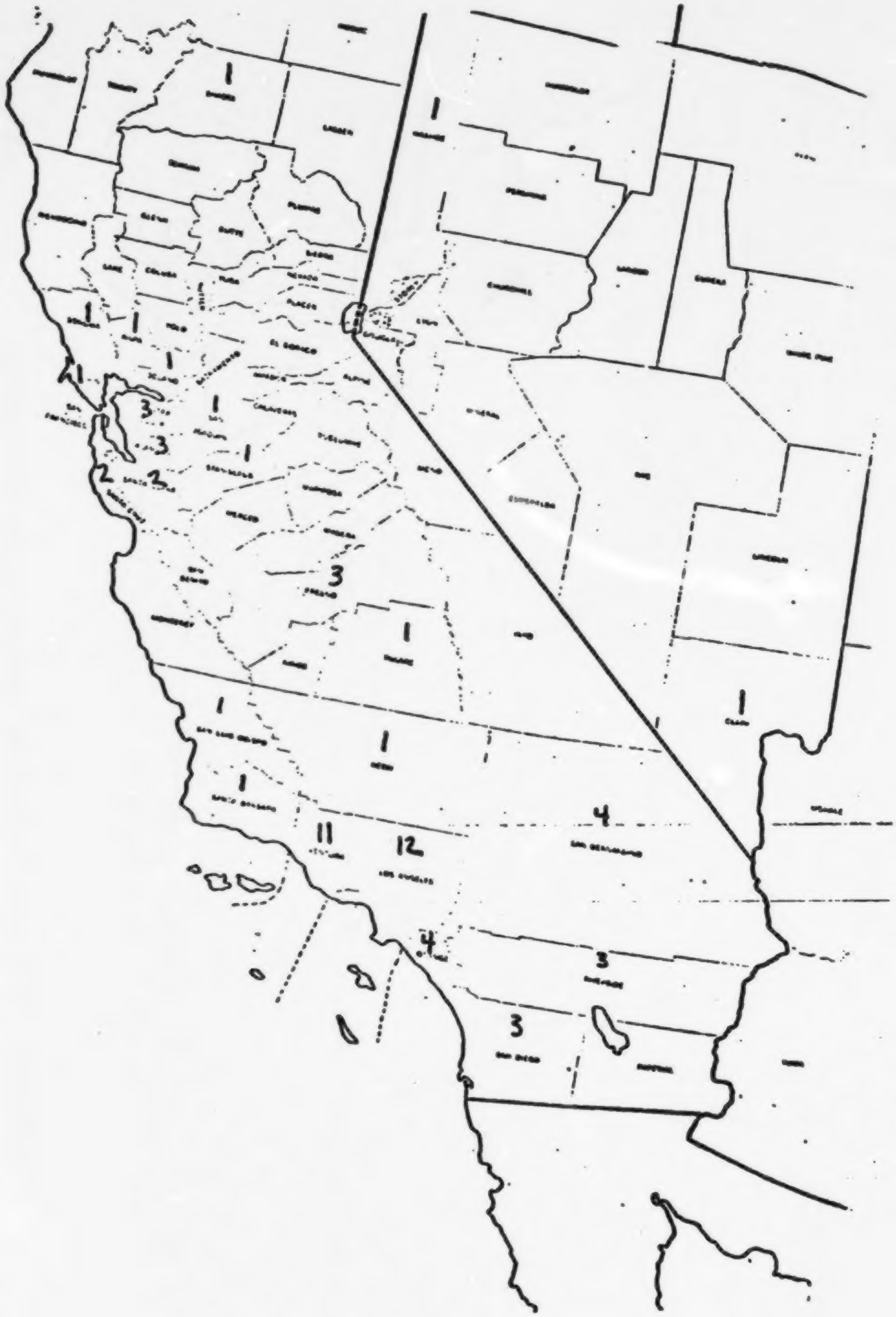

1
1
1
1
1
1
3
1
3
1
2
2
3
1
1
1
1
1
4
11
12
4
3
3

Figure C-1. 1973 CO Monitoring Station Sample Distribution

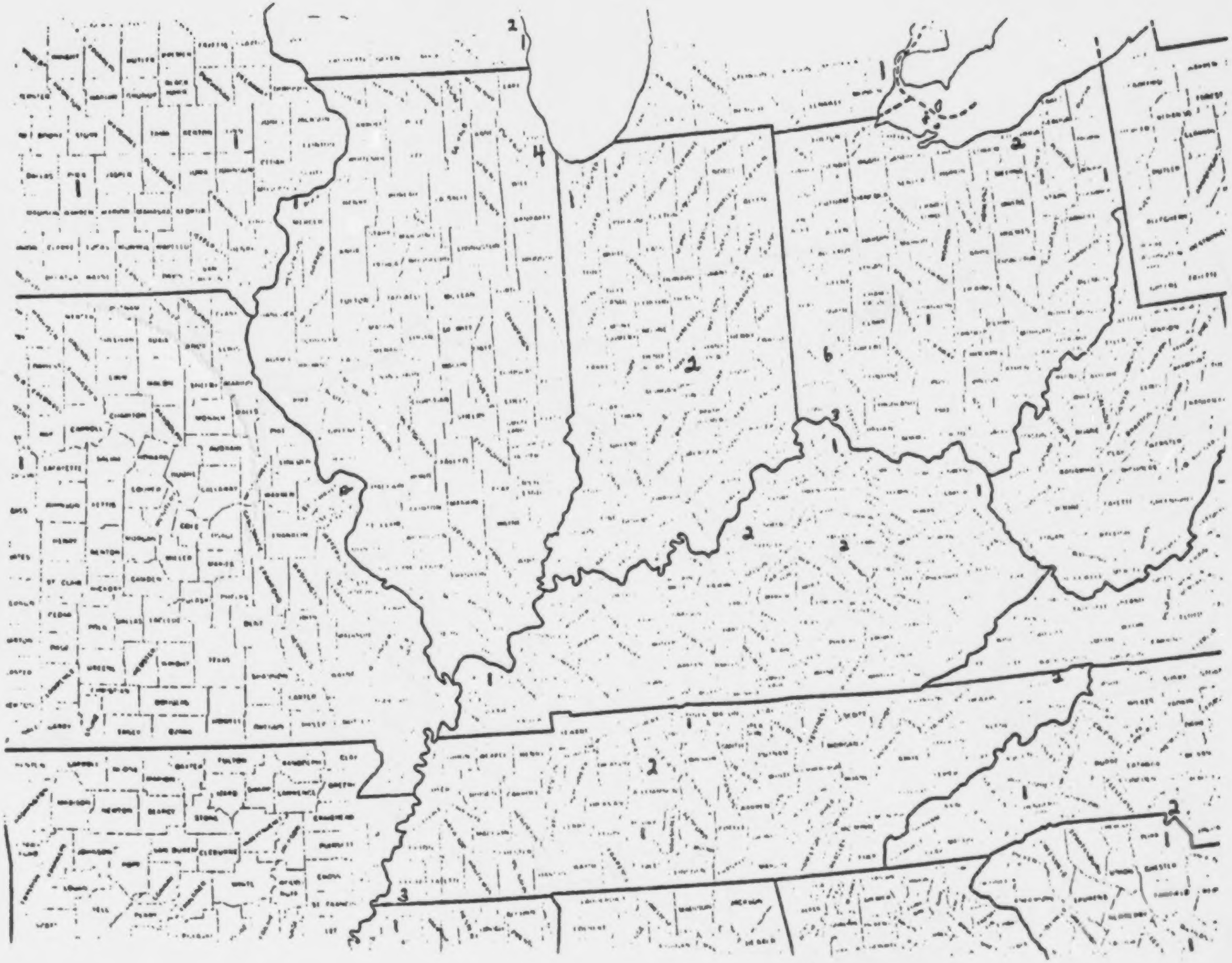


Figure C-1. 1973 Ozone Monitoring Station Sample Distribution

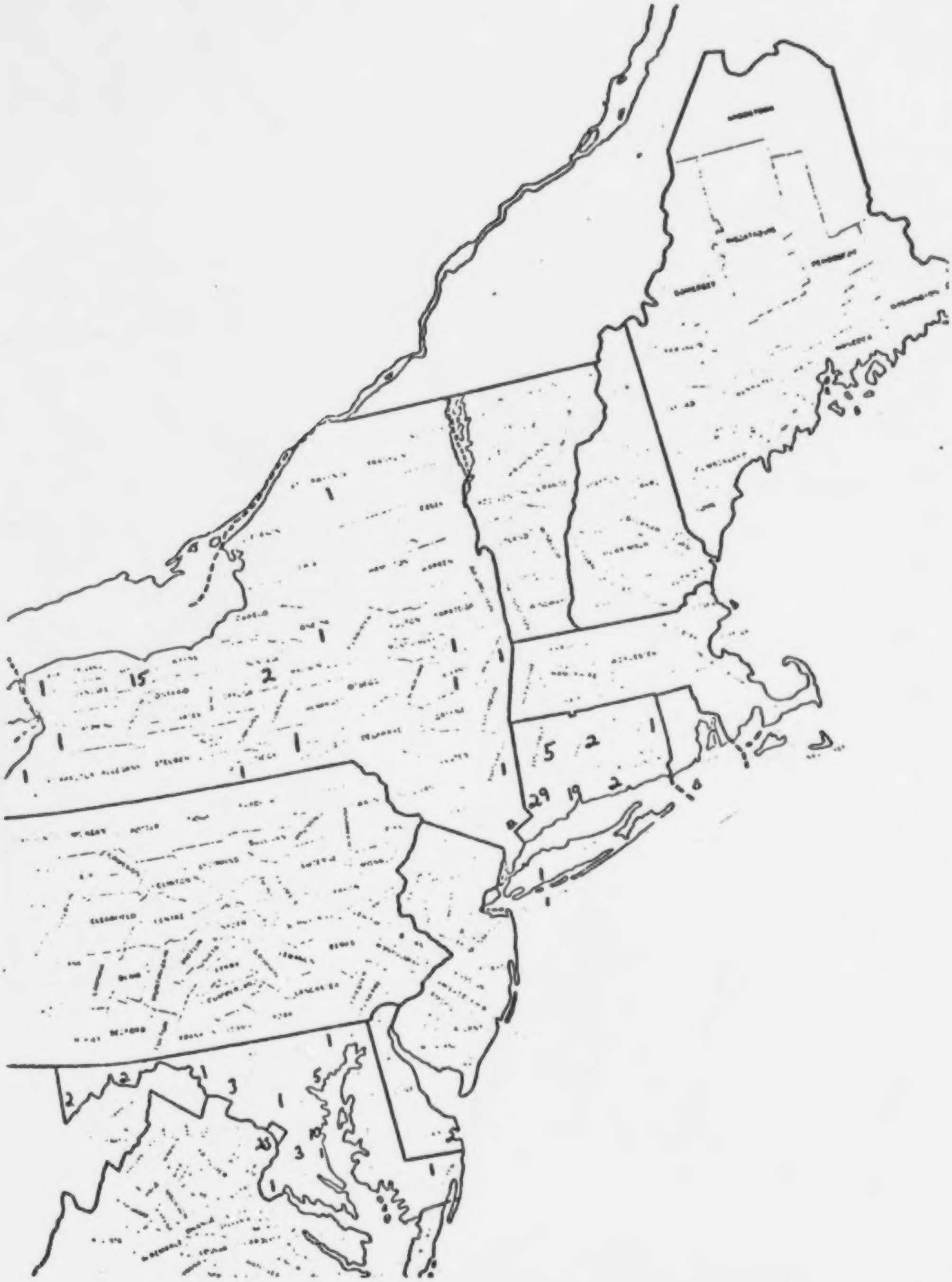
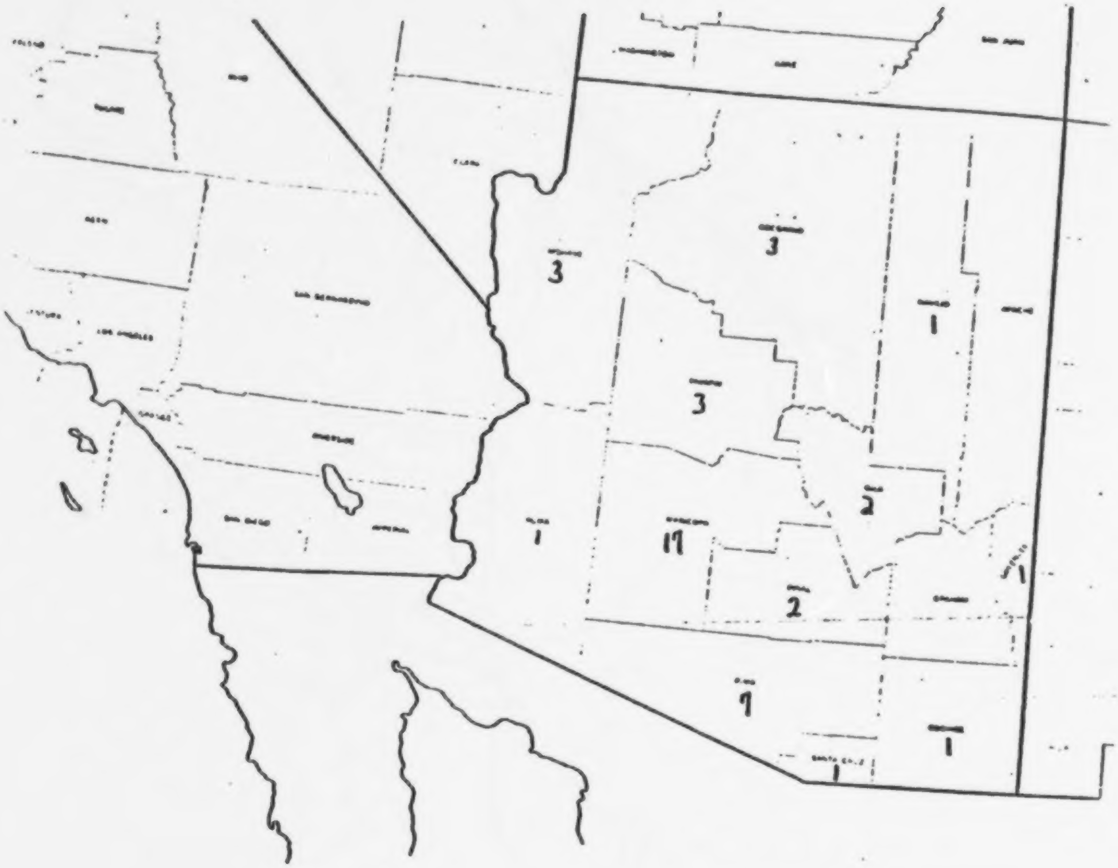

1
1
1
1
15
2
1
1
1
1
1
1
5
2
1
29
19
2
1
1
2
2
1
3
1
5
20
3
10
1
1
1

Figure C-1. 1973 Sulfates Monitoring
Station Sample Distribution


3
3
1
3
2
1
17
1
2
9
SANTA CRUZ
1
1

Reproduced from best available copy.

Figure C-1. 1973 Nitrates Monitoring Station Sample Distribution

TABLE C-2. PERCENTAGE OF STATES' POPULATION IN COUNTIES WITH ONE OR MORE MONITORING STATIONS

1973 TSP

State	at least 1	at least 2	at least 3	at least 5	at least 10
AL	77	52	48	43	25
AZ	97	90	84	75	55
AR	61	36	27	3	0
CA	93	82	71	35	35
CO	93	87	73	27	0
CT	92	92	89	78	78
DE	100	100	70	70	70
FL	81	71	62	51	27
GA	56	37	37	17	17
ID	58	58	23	16	0
IL	84	77	71	65	54
IN	66	47	45	42	33
IA	53	35	22	16	10
KS	77	43	41	41	0
KY	33	18	14	11	0
LA	43	31	0	0	0
ME	74	56	43	32	0
MD	81	72	67	63	53
MA	90	80	79	71	36
MI	85	78	69	66	47
MN	69	58	54	49	44
MS	54	22	16	16	0
MO	71	63	54	54	48
MT	60	42	32	27	0
NB	67	45	38	38	11
NV	98	86	83	81	81
NH	100	97	76	16	0
NJ	84	78	77	49	8
NM	82	66	56	51	0
NY	100	94	90	83	48
NC	70	47	39	29	11
ND	59	38	12	0	0
OH	69	61	58	51	42
OK	92	61	45	36	21
OR	94	74	49	45	27
PA	81	78	67	56	30
RI	100	95	80	70	61
SC	85	51	49	37	25
SD	31	23	14	0	0
TN	80	68	53	47	36
TX	71	60	55	55	52
UT	68	55	43	43	0
VT	62	12	0	0	0
VA	82	63	44	27	17
WA	95	77	61	54	34
WV	46	46	34	17	13
WI	73	60	49	44	28
WY	67	30	0	0	0
Total US	79	66	58	49	34

TABLE C-2. PERCENTAGE OF STATES' POPULATIONS IN COUNTIES WITH ONE OR MORE MONITORING STATIONS

1973 SO_2

State	at least 1	at least 2	at least 3	at least 5	at least 10
AL	37	27	19	19	0
AZ	84	83	80	80	20
AR	25	20	17	0	0
CA	80	69	52	45	35
CO	42	23	23	23	0
CT	100	93	93	90	78
DE	85	70	70	70	70
FL	64	58	34	10	9
GA	36	21	17	17	13
ID	3	3	3	3	0
IL	67	58	53	49	49
IN	61	51	39	32	26
IA	46	21	13	10	0
KS	61	46	41	31	16
KY	72	54	44	41	27
LA	55	44	44	0	0
ME	74	66	53	32	0
MD	74	67	58	56	37
MA	88	79	79	71	36
MI	67	59	48	40	30
MN	59	55	49	44	25
MS	36	24	14	14	4
MO	43	37	35	35	20
MT	28	15	3	2	0
NB	38	38	26	26	0
NV	2	2	2	2	0
NH	61	35	35	30	0
NJ	90	67	19	6	0
NM	58	55	46	14	5
NY	62	51	45	24	19
NC	89	62	38	29	11
ND	0	0	0	0	0
OH	61	54	49	35	25
OK	55	43	38	21	21
OR	57	27	27	0	0
PA	50	33	33	33	0
RI	95	85	70	70	61
SC	78	54	47	38	10
SD	0	0	0	0	0
TN	59	54	51	44	18
TX	68	57	56	45	16
UT	68	43	43	43	43
VT	45	0	0	0	0
VA	56	38	23	10	0
WA	81	66	58	0	0
WV	25	19	15	13	0
WI	59	44	36	34	0
WY	43	15	0	0	0
Total US	64	52	43	33	19

TABLE C-2. PERCENTAGE OF STATES' POPULATION IN COUNTIES WITH ONE OR MORE MONITORING STATIONS

1973 NO$_2$

State	at least 1	at least 2	at least 3	at least 5	at least 10
AL	35	26	21	21	0
AZ	79	75	0	0	0
AR	17	15	0	0	0
CA	96	82	77	60	35
CO	23	23	23	23	0
CT	100	93	90	82	26
DE	85	70	0	0	0
FL	47	34	27	19	0
GA	21	21	13	13	13
ID	0	0	0	0	0
IL	53	53	49	49	49
IN	56	42	27	11	11
IA	16	13	10	10	0
KS	43	31	7	0	0
KY	72	52	44	38	23
LA	30	30	0	0	0
ME	23	4	0	0	0
MD	74	64	58	56	16
MA	52	0	0	0	0
MI	48	45	40	0	0
MN	52	47	44	38	25
MS	14	14	0	0	0
MO	37	37	34	34	20
MT	2	0	0	0	0
NB	38	11	0	0	0
NV	81	0	0	0	0
NH	5	0	0	0	0
NJ	50	49	35	0	0
NM	36	31	0	0	0
NY	48	24	15	0	0
NC	24	20	3	3	0
ND	0	0	0	0	0
OH	55	54	23	9	0
OK	37	21	21	0	0
OR	27	27	0	0	0
PA	50	20	17	17	17
RI	95	70	70	61	61
SC	79	47	45	38	0
SD	1	1	1	1	0
TN	39	30	0	0	0
TX	61	51	21	16	0
UT	43	0	0	0	0
VT	0	0	0	0	0
VA	19	0	0	0	0
WA	54	34	0	0	0
WV	17	0	0	0	0
WI	44	7	0	0	0
WY	21	0	0	0	0
Total US	51	38	27	18	10

TABLE C-2. PERCENTAGE OF STATES' POPULATION IN COUNTIES WITH ONE OR MORE MONITORING STATIONS

1973 CO

State	at least 1	at least 2	at least 3	at least 5	at least 10
AL	10	0	0	0	0
AZ	75	75	0	0	0
AR	0	0	0	0	0
CA	93	75	61	38	35
CO	42	23	23	23	0
CT	78	26	0	0	0
DE	0	0	0	0	0
FL	24	19	0	0	0
GA	13	13	0	0	0
ID	0	0	0	0	0
IL	52	49	49	49	0
IN	15	0	0	0	0
IA	16	0	0	0	0
KS	46	24	0	0	0
KY	5	1	1	0	0
LA	16	0	0	0	0
ME	0	0	0	0	0
MD	66	61	44	23	23
MA	70	46	38	0	0
MI	39	30	30	0	0
MN	6	0	0	0	0
MS	0	0	0	0	0
MO	50	34	34	34	0
MT	0	0	0	0	0
NB	38	0	0	0	0
NV	81	0	0	0	0
NH	30	30	0	0	0
NJ	94	44	22	0	0
NM	43	43	43	36	5
NY	31	21	8	8	0
NC	3	0	0	0	0
ND	0	0	0	0	0
OH	36	27	16	0	0
OK	36	21	21	0	0
OR	37	27	0	0	0
PA	30	30	17	0	0
RI	61	61	61	61	0
SC	3	0	0	0	0
SD	0	0	0	0	0
TN	33	30	18	0	0
TX	42	0	0	0	0
UT	68	43	0	0	0
VT	22	0	0	0	0
VA	32	10	10	0	0
WA	64	42	42	0	0
WV	13	0	0	0	0
WI	18	24	24	0	0
WY	0	0	0	0	0
Total US	41	27	20	11	6

TABLE C-2. PERCENTAGE OF STATES' POPULATIONS IN COUNTIES WITH ONE OR MORE MONITORING STATIONS

1973 OZONE

State	at least 1	at least 2	at least 3	at least 5	at least 10
AL	0	0	0	0	0
AZ	75	0	0	0	0
AR	0	0	0	0	0
CA	96	83	73	50	35
CO	42	23	23	23	0
CT	82	26	26	26	0
DE	0	0	0	0	0
FL	41	27	19	0	0
GA	13	0	0	0	0
ID	0	0	0	0	0
IL	51	49	49	0	0
IN	26	15	0	0	0
IA	16	0	0	0	0
KS	33	24	0	0	0
KY	33	27	0	0	0
LA	28	16	0	0	0
ME	0	0	0	0	0
MD	46	46	29	0	0
MA	59	38	25	0	0
MI	30	0	0	0	0
MN	38	0	0	0	0
MS	15	14	0	0	0
MO	37	20	20	20	20
MT	0	0	0	0	0
NB	26	0	0	0	0
NV	56	56	0	0	0
NH	35	0	0	0	0
NJ	39	15	0	0	0
NM	31	31	31	31	0
NY	31	9	0	0	0
NC	10	7	0	0	0
ND	0	0	0	0	0
OH	46	31	14	6	0
OK	36	21	0	0	0
OR	34	27	0	0	0
PA	17	17	17	17	0
RI	61	61	0	0	0
SC	9	0	0	0	0
SD	0	0	0	0	0
TN	35	33	18	0	0
TX	22	14	0	0	0
UT	68	43	0	0	0
VT	0	0	0	0	0
VA	23	5	5	0	0
WA	34	34	34	34	0
WV	0	0	0	0	0
WI	24	23	0	0	0
WY	0	0	0	0	0
Total US	37	25	16	8	4

TABLE C-2. PERCENTAGE OF STATES' POPULATIONS IN COUNTIES WITH ONE OR MORE MONITORING STATIONS

1973 SULFATES

State	at least 1	at least 2	at least 3	at least 5	at least 10
AL	0	0	0	0	0
AZ	97	86	78	75	55
AR	0	0	0	0	0
CA	0	0	0	0	0
CO	0	0	0	0	0
CT	90	86	56	56	51
DE	0	0	0	0	0
FL	0	0	0	0	0
GA	60	19	8	0	0
ID	0	0	0	0	0
IL	0	0	0	0	0
IN	42	19	12	12	0
IA	0	0	0	0	0
KS	0	0	0	0	0
KY	24	12	5	5	2
LA	0	0	0	0	0
ME	0	0	0	0	0
MD	55	45	43	23	8
MA	0	0	0	0	0
MI	0	0	0	0	0
MN	69	59	56	49	44
MS	0	0	0	0	0
MO	21	17	14	14	0
MT	8	8	8	0	0
NB	0	0	0	0	0
NV	25	25	25	25	25
NH	0	0	0	0	0
NJ	0	0	0	0	0
NM	0	0	0	0	0
NY	44	11	4	4	4
NC	0	0	0	0	0
ND	59	59	31	12	0
OH	9	9	9	9	9
OK	0	0	0	0	0
OR	0	0	0	0	0
PA	0	0	0	0	0
RI	0	0	0	0	0
SC	0	0	0	0	0
SD	0	0	0	0	0
TN	7	3	0	0	0
TX	17	17	17	17	17
UT	0	0	0	0	0
VT	0	0	0	0	0
VA	0	0	0	0	0
WA	0	0	0	0	0
WV	0	0	0	0	0
WI	0	0	0	0	0
WY	0	0	0	0	0
Total US	14	8	6	5	4

TABLE C-2. PERCENTAGE OF STATES' POPULATIONS IN COUNTIES WITH ONE OR MORE MONITORING STATIONS

1973 NITRATES

State	at least 1	at least 2	at least 3	at least 5	at least 10
AL	9	0	0	0	0
AZ	97	86	81	75	55
AR	0	0	0	0	0
CA	0	0	0	0	0
CO	0	0	0	0	0
CT	0	0	0	0	0
DE	0	0	0	0	0
FL	0	0	0	0	0
GA	0	0	0	0	0
ID	0	0	0	0	0
IL	0	0	0	0	0
IN	0	0	0	0	0
IA	0	0	0	0	0
KS	0	0	0	0	0
KY	0	0	0	0	0
LA	0	0	0	0	0
ME	0	0	0	0	0
MD	0	0	0	0	0
MA	0	0	0	0	0
MI	0	0	0	0	0
MN	0	0	0	0	0
MS	0	0	0	0	0
MO	0	0	0	0	0
MT	0	0	0	0	0
NB	0	0	0	0	0
NV	0	0	0	0	0
NH	0	0	0	0	0
NJ	0	0	0	0	0
NM	0	0	0	0	0
NY	44	11	4	4	4
NC	0	0	0	0	0
ND	59	31	12	0	0
OH	0	0	0	0	0
OK	0	0	0	0	0
OR	0	0	0	0	0
PA	0	0	0	0	0
RI	0	0	0	0	0
SC	0	0	0	0	0
SD	0	0	0	0	0
TN	0	0	0	0	0
TX	0	0	0	0	0
UT	0	0	0	0	0
VT	0	0	0	0	0
VA	0	0	0	0	0
WA	0	0	0	0	0
WV	0	0	0	0	0
WI	0	0	0	0	0
WY	0	0	0	0	0
Total US	5	2	1	1	9

CORRELATION AND REGRESSION STUDIES: TSP-NEDS EMISSIONS

BACKGROUND

Results of the survey of monitoring activity show that 79% of the U.S. population (1970 Census data) resided in counties having at least one station monitoring TSP in 1973. Even if all stations reported valid data there were nevertheless more than 42 million Americans whose resident counties were not being monitored for TSP. Instrumental techniques for monitoring this pollutant are well-established and stable so that one would expect that the ambient monitoring data are reliable. It was therefore decided to undertake a study to investigate possible correlations between TSP and various emissions and meteorological parameters on a county basis. The results of these studies were used to augment the air quality data for counties with no monitoring for TSP but which had emmissions for TSP and NO_x lying within the range of the regression analysis.

METHODOLOGY AND RESULTS

A file of mean annual ambient air quality for all stations reporting valid TSP data to SAROAD in 1973 was assembled. Although exploratory studies were made using all such data, the final work reported here was performed using only data from stations sited to provide ambient air quality measurements for the resident population. Table C-3 lists the numbers of counties in various categories referring to level of monitoring activity in those individual counties. Thus the entry '114' under category '$N \geq 4$' means that a total of 114 counties had 4 or more stations reporting valid yearly SAROAD readings for TSP. The '$N \geq 1$' category contains the entire data set.

The division of valid stations into these categories was made for the following reasons:

1. It was intended to develop a regression model relating mean annual TSP ambient air quality to NEDS emission values independently for data in a number of these categories. The detailed forms of the models thus obtained could be compared and the model giving the best 'over all' fit selected. It is, of course, essential that the functional form of the

regression model does not change as more or fewer stations, selected on some rational basis, are included in the data set.

2. The model selected as a result of the above considerations must then be tested against a known data set. This data set could come only from the valid SAROAD readings. The procedure followed was to derive the model from a comparison of results in the N≥4, N≥6, N≥7, N≥10 subgroups and then to compare the predicted and measured values for the entire 605 point data set.

3. It is of some interest to determine the maximum correlation coefficient obtainable as the N categories are scanned since this result has implications for determining how representative a given station's monitoring is of emissions in the county.

For each county reporting some valid yearly data to SAROAD, both the geometric and arithmetic means of the annual average for all stations reporting valid data within the county were computed. Since air pollution data tends to be lognormally distributed, the geometric means of the annual averages will also be lognormally distributed if one assumes that the ambient air quality annual averages at stations within a county are independently distributed. Although this is not likely to be a good assumption, it was felt important to study correlations of the geometric means as well as the arithmetic means because of the simplicity of the resulting distribution. Actually, the results of the study showed no significant difference in the structure of the models resulting from the use of either of these alternative averaging procedures. The magnitudes of the correlation coefficients were likewise similar.

Standard correlation and regression analyses were run for a variety of empirical models using the SAS (Statistical Analysis System) available on EPA's UNIVAC 1110. After experience with numerous trials the following requirements for an acceptable model were adopted:

1. A correlation coefficient, R, of 0.7 or greater,
2. Parsimony and simplicity of the resulting model,
3. Stability of model parameters with respect to variation of the N category of the data set.

Table C-4 shows the maximum correlation values obtained for the four N categories used as a function of the number of parameters used in the regression model. The possible parameters in the model are listed at the top of the table and the resulting parameter set which maximized R are listed in the table proper. Note that 'TSP', which is the NEDS total TSP county emission value, makes the major contribution to explaining the data variance in all cases. Significant contributions are made also by high powers of NO_x and HC NEDS county total emissions. It should also be noted that the maximum value for R obtainable with this parameter set decreases monotonically with the value of the N category.

The model chosen as giving the best 'overall fit' as a result of many trials and comparison of maximum and near maximum values for R was a linear regression model having NEDS county total TSP, $(NO_x)^4$, and $(NO_x)^5$ as the independent variables.

Table C-5 gives a comparison between the maximum correlation coefficients from Table 2 for a three parameter model and the R values obtained from the chosen theory for the four N categories used. Except for the N≥10 categories differences in R are in the third decimal place.

Table C-6 gives a similar comparison, for the N≥7 category between the chosen model and a sampling of other 'simple' models. Here again the differences are small.

While it is our judgement that the model chosen gives the best overall fit to the data in the N≥4, N≥6, N≥7 and N≥10 categories the differences in fit between it and alternative models are small enough so that it does not seem appropriate to perform any attempts at theoretically justifying the resulting functional form. Rather the model should be viewed as a purely empirical means of calculating TSP air quality in counties for which data is missing or inadequate. In this regard, it should be noted that an attempt was made to impose the theoretically justifiable functional form $\frac{TSP}{U}$ (where U is the county mean annual wind speed) on the regression model with disappointing results.

While the model should be viewed as purely empirical, it is nevertheless important to ensure that the choice of high powers of the NO_x NEDS county total emissions is not the result of a few outlying data points. Accordingly, 5 points of the 58 point data set for the N≥7 category were chosen for removal from the data set. The choice was made as follows:

1. A single parameter model using only NEDS county total TSP emissions was used to predict air quality in the 58 counties of interest. The 5 counties giving the worst agreement with measured averages were then removed from the data set and the resulting 53 point data set was reanalyzed. Table C-6 gives the result of this analysis by showing a comparison between the 1st and 2nd best values for R for the two data sets along with the resulting variable sets. It can be seen that the model chosen holds up quite well, differences in R being mainly in the third decimal place.

Having chosen the model, we calculated the TSP ambient air quality for all 605 counties for which valid data was available using regression coefficients derived from the N≥7 category data. The ratio of measured to calculated values were then computed and the results plotted on the frequency graph Figure C-2. In this figure, the computed ratios appear as the discontinuous curves labelled 'data' while the smooth curve is a fitted normal distribution. It can be seen from this figure that the empirical theory developed predicts the TSP yearly average ambient air quality within about ±37% for a 67% confidence interval. The population variance σ^2 and its associated 95% confidence interval are $\sigma_1^2 = .076 \pm .009$.

Part of the variance of the prediction is attributable to an 'inherent' variability in the data which results from the fact that within a given county measured air quality at monitoring stations itself varies. Figure C-3 gives a measure of this variability. The ratio of individual station data within a county to geometric mean for all stations within that county were computed and plotted on this figure. The variance of the resulting fitted normal curve and its associated 95% confidence interval is $\sigma_2^2 = .056 \pm .004$.

Figure C-4 shows the two fitted normal curves from Figures C-2 and C-3 plotted to the same scale and superimposed. Comparison of σ_1^2 and σ_2^2 indicates that not much improvement in the predictive value of the regression model can be obtained. We are, however, currently investigating the use of $\exp(\alpha NO_x)$ as a possible parameter primarily because of the simplicity of its functional form.

Figure C-5 shows the variation of the square of the correlation coefficient, R^2, for the chosen model with the number of points included in the correlation data set. The numbers on the curve indicate the corresponding N category. Provided that errors in NEDS emission data show no systematic correlation with N category (which seems a reasonable assumption) the monotonic decrease of R^2 with N value indicates that representativeness of population based monitoring requires something in the neighborhood of 7 or more stations per county with present siting strategies. Thus, for example, if one were constrained to a fixed number of monitoring stations nationwide and/or if one were interested in reducing the amount of monitoring with little or no reduction in the overall quality of monitoring data, the results presented here are highly suggestive of a strategy which would consist of eliminating monitoring in most or all counties having only a few stations monitoring population ambients and instead concentrating one's effort to place at least seven stations in a smaller number of counties representative of a wide range of NEDS emissions for TSP and NO_x nationwide.

In terms of the primary aim of this contract, i.e., to develop a data base including air quality measures for all counties in the contiguous United States, the regression model presented in this report would seem quite adequate for filling in some missing TSP data.

TABLE C-3. CUMULATIVE FREQUENCY TABLES FOR VALID 1973 SAROAD POPULATION BASED MONITORING OF TSP

Category N: Number of Stations in County Reporting Valid Population Based TSP Data to SAROAD in 1973	$N \geq 10$	$N \geq 9$	$N \geq 8$	$N \geq 7$	$N \geq 6$	$N \geq 5$	$N \geq 4$	$N \geq 3$	$N \geq 2$	$N \geq 1$
Total Number of Stations in Given 'N' category (number of points in given data set)	27	33	42	58	71	87	114	170	266	605

TABLE C-4. MAXIMUM CORRELATION VALUES, R, FOR A SAMPLING OF DIFFERENT MODELS INVESTIGATED

Possible Emission Variables in Model: $(TSP)^i$; $(SO_2)^j$; $(NO_x)^k$; $(HC)^\ell$; $(CO)^m$
where $i, j, k, \ell, m = 0, 1, 2, 3, 4, 5$

N Category	Number of Parameters Needed	Variables in model which maximize R	Resulting Correlation Coefficient T
N≥4	1	TSP	.43
	2	TSP, $(NO_x)^5$	.46
	3	TSP, $(NO_x)^4$, $(NO_x)^5$	.49
	4	TSP, $(NO_x)^4$, $(NO_x)^5$,CO	.51
N≥6	1	TSP	.50
	2	TSP, $(HC)^5$	.55
	3	TSP, $(NO_x)^2$ $(HC)^5$	.59
	4	TSP, $(NO_x)^2$, $(HC)^5$, CO	.61
N≥7	1	TSP	.60
	2	TSP, $(HC)^5$	.66
	3	TSP, $(NO_x)^2$, HC^5	.71
	4	TSP, $(SO_2)^2$, $(NO_x)^2$, $(HC)^3$	.73
N≥10	1	TSP	.67
	2	$(TSP)^2$, $(NO_x)^5$	.78
	3	$(TSP)^4$, SO_2, $(NO_x)^5$	.84
	4	$(TSP)^5$, SO_2, $(NO_x)^4$, $(NO_x)^5$	.84

TABLE C-5. SAMPLE COMPARISON BETWEEN MAXIMUM CORRELATION COEFFICIENTS R AND THOSE OBTAINED FROM MODEL CHOSEN AS GIVING BEST OVERALL FIT (THREE PARAMETER THEORY)

Chosen model: TSP; NO_x^4; NO_x^5

N Category	Variables which maximize R	Maximum R	Number of points in data set	R for Model Chosen
N≥4	TSP, NO_x^4, NO_x^5	.49	114	.49
N≥6	TSP, NO_x^2, HC^5	.59	71	.59
N≥7	TSP, NO_x^2, HC^5	.71	58	.71
N≥10	TSP^4, SO_2, NO_x^5	.84	27	.77

TABLE C-6. SAMPLE COMPARISON BETWEEN CHOSEN MODEL AND OTHER 'SIMPLE' MODELS ($N \geq 7$)

	Variables	R	
3 parameter	TSP, NO_x^4, NO_x^5	.71	Chosen model
	TSP NO_x^3, NO_x^5	.70	
	TSP, NO_x^3, NO_x^4	.70	
	TSP, NO_x^2, e^{NO_x}	.70	
2 parameter	TSP, NO_x^5	.64	
	TSP, NO_x^4	.63	
	TSP, e^{NO_x}	.63	

TABLE C-7. EFFECT ON REGRESSION MODEL OF REMOVING 5 'OUTLIER' POINTS FROM 58 POINT DATA SET BASED ON SINGLE PARAMETER THEORY

53 Point Data Set			Original 58 Point Data Set		
Number of Parameters	1st & 2nd Best Variable Set	R	Number of Parameters	1st & 2nd Best Variable Set	R
1	TSP	.626	1	TSP	.598
1	NO_x	.628	1	$(NO_x)^2$	.396
2	TSP, $(NO_x)^4$	.653	2	TSP, $(NO_x)^5$	.642
2	TSP, $(NO_x)^5$	.653	2	TSP, $(NO_x)^4$	.643
3	TSP, $(NO_x)^2$, $(NO_x)^3$	.654	3	TSP, $(NO_x)^4$, $(NO_x)^5$	.707
3	TSP, $(NO_x)^4$, $(NO_x)^5$	.653	3	TSP, $(NO_x)^3$, $(NO_x)^5$	.705

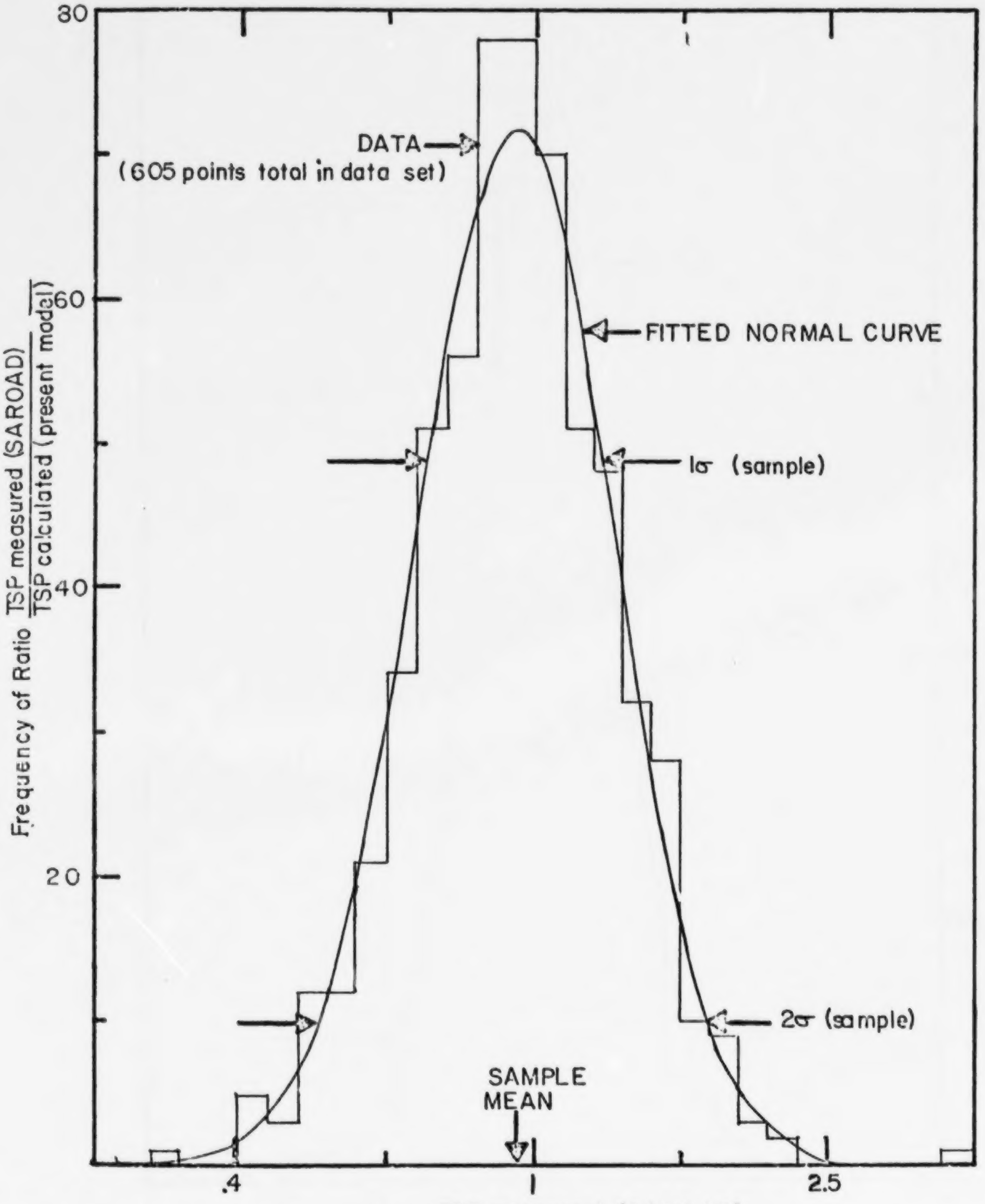


FIG. C-2. Fit of Present Model with all Valid 1973 SAROAD Population Based Monitoring

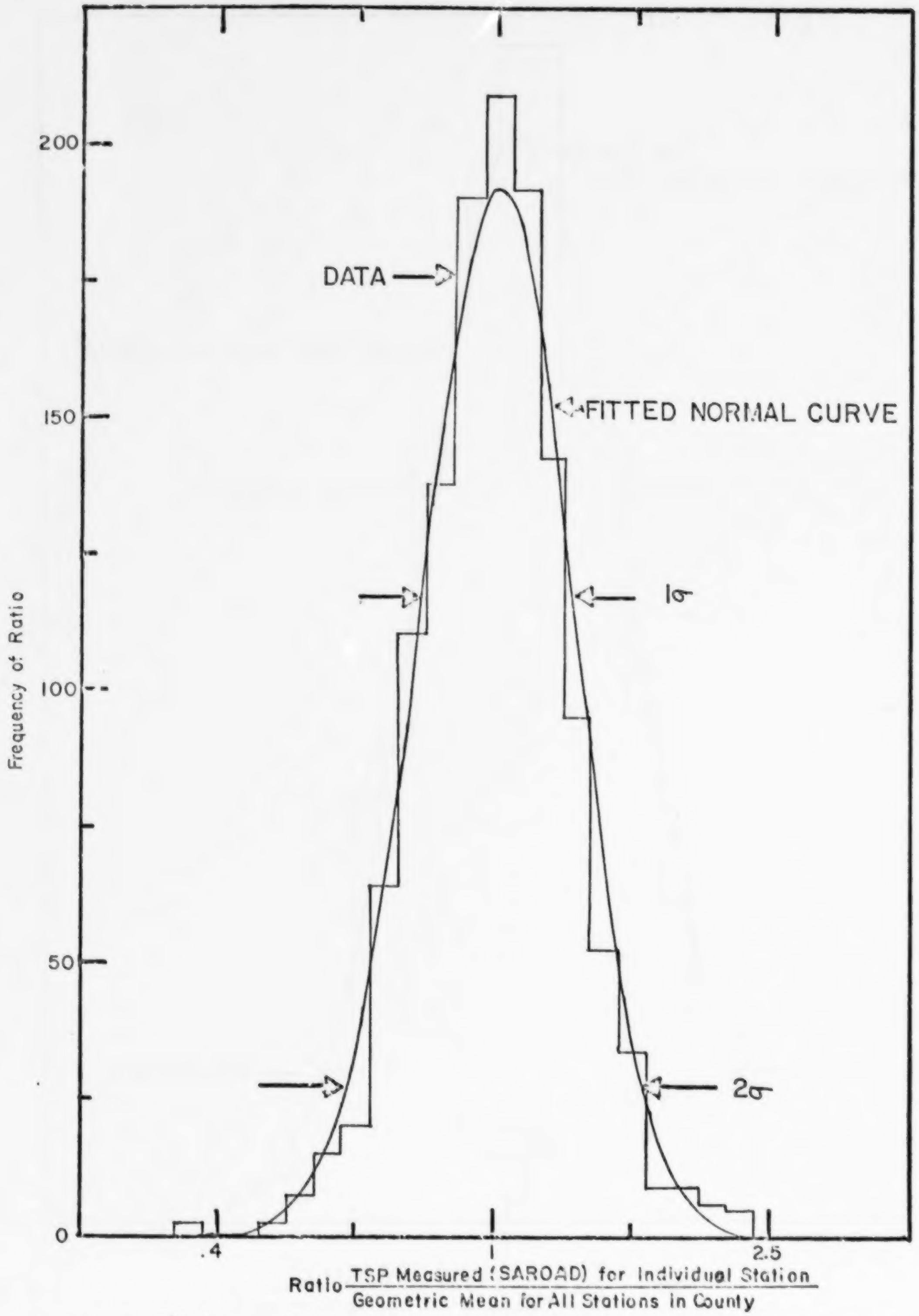


FIG. C-3. 'Inherent' Variability of SAROAD Data for Intracounty Population Based Monitoring

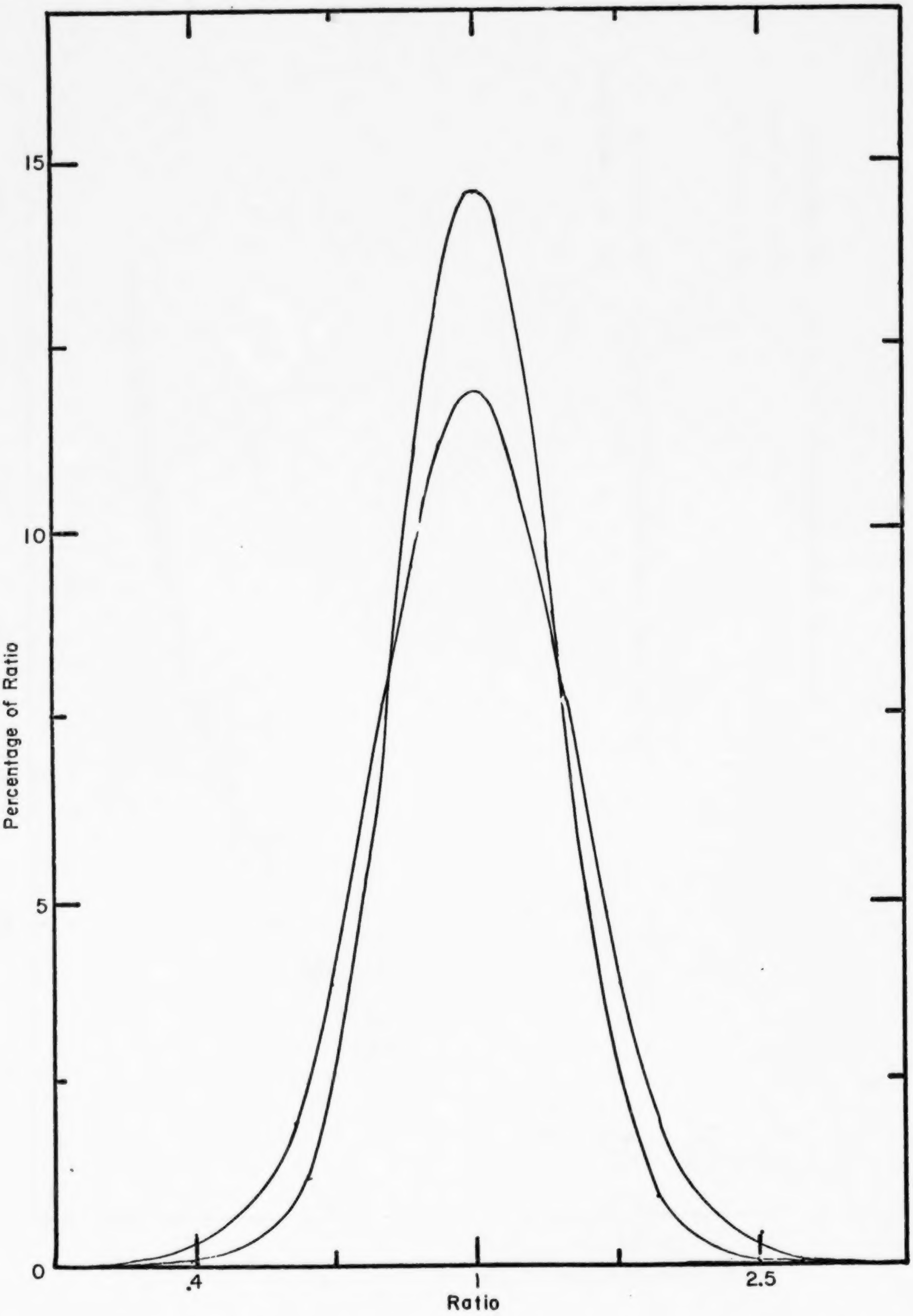


FIG. C-4. Comparison of Fitted Normal Curve for Predicted Air Quality and 'Inherent' Variability

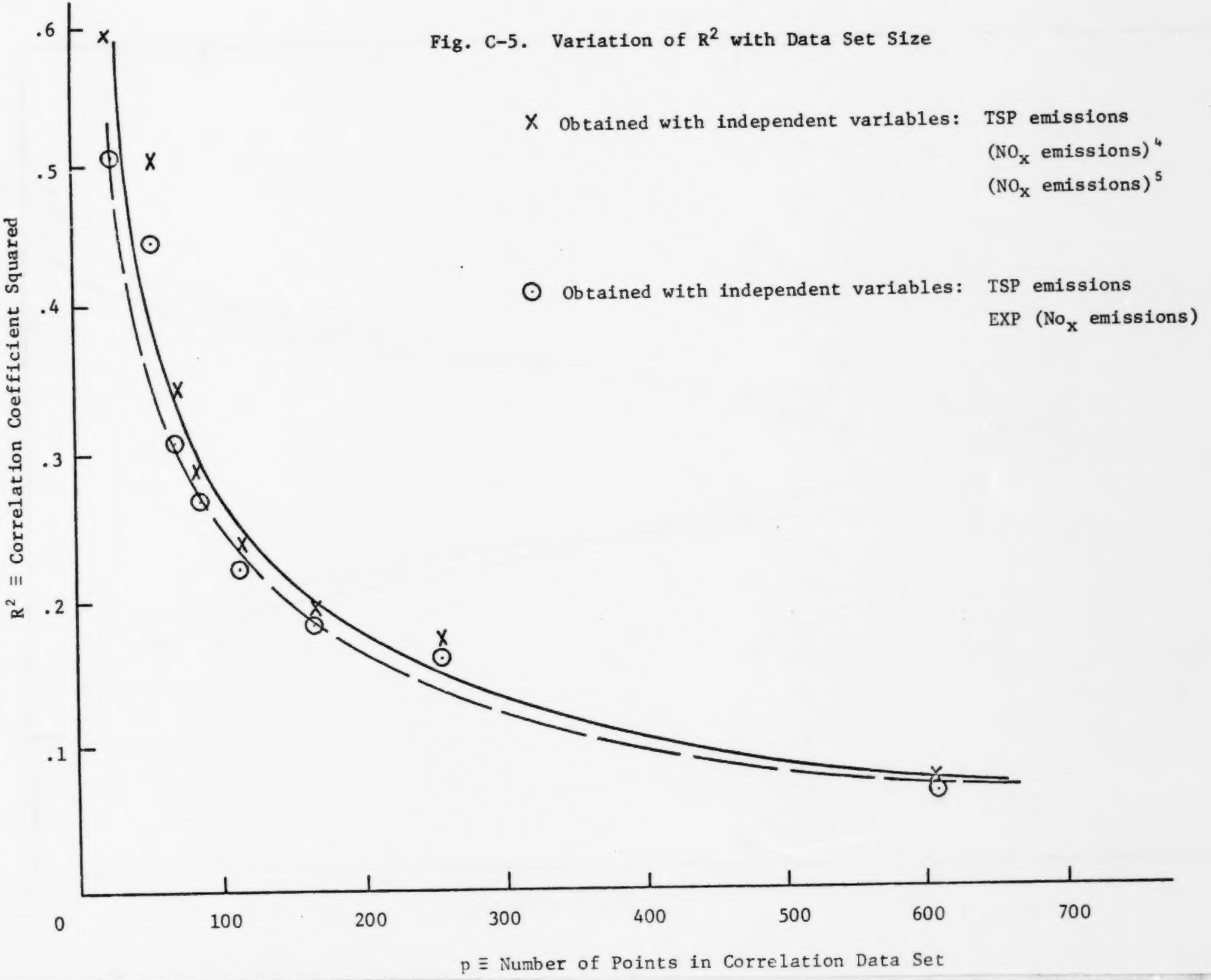


Fig. C-5. Variation of R^2 with Data Set Size

CORRELATION AND REGRESSION STUDIES: SO_2 - GROUND DEPOSITION

A rigorous determination of population-at-risk with respect to air polution exposure requires a prohibitive level of air monitoring to account for local variations. Many counties lack the financial resources and technical expertise to successfully operate a sufficient number of monitoring stations. Therefore, the population-at-risk study inevitably involves some extrapolation of a limited air quality data set. The errors introduced by the extrapolation represent a limitation to the accuracy of the risk estimates. Selected factors which could conceivably introduce local variations in air quality are discussed below followed by an attempt to incorporate one of the dominant factors into the statistical models used to predict air pollution exposure.

POLLUTANT REACTIVITY AND ATMOSPHERIC "CLEANSING" PROCESSES

Analyses discussed above consider only the SO_2 emission density and the general source characterization (point and area). While these are probably the dominant factors influencing the local SO_2 levels on an annual average, there are a variety of additional factors which can affect observed air quality. A partial list is as follows:

- Ground Deposition Processes
- Regional Transport (Inter-county dispersion)
- Atmospheric reactivity (smog chemistry) of SO_2
- SO_2 Washout and Rainout
- Plume reactivity of SO_2
- Type of atmospheric monitors
- Monitoring site characteristics
- Monitor operational characteristics
- Micrometeorological factors

The latter group of factors has a potentially large impact on observed air quality; however, evaluation requires a time-consuming, comprehensive study of local conditions which is beyond the scope of the present project. Therefore, attempts to improve the prediction capabilities of the models are limited to consideration of atmospheric reactivity and pollutant removal processes.

SO_2 reactivity in large point source plumes has been observed to range from 2 to 15 percent per hour depending on meteorological conditions and measuring technique.[1] On an assumed average basis this does not appear to be a major factor; particularly since area sources of SO_2 are dominant in most regions. Area sources emit SO_2 at levels subject to surface air flow disturbances and therefore, experience less defined plumes.

Washout and rainout are one of the two major atmospheric removal processes; however, conditions are relatively similar in a given region of the county to adjacent counties in the same general region.

Atmospheric reactivity, regional transport and ground deposition processes appear to be the dominant factors potentially introducing local variation in SO_2 levels. The relative importance of these has been evaluated by Richards and Gerstle[1] using a Continous Stirred Tank Reactor (CSTR) model. This approach involves CSTR equations for a first order reaction process in conjunction with NEDS emission data and SAROAD ambient SO_2 data in a 10 state northwestern region. Comparison of the predicted SO_2 annual average concentration at various decay rates with observed concentrations are used to identify the prevailing SO_2 decay rate and the SO_2 oxidation rate. By dividing the 10 state west coast region into four approximately equal regions, it was possible to roughly calculate the extent of regional transport.

The results of the CSTR calculations suggest that atmospheric reactivity is relatively insignificant compared with ground deposition and other "cleansing processes." Most SO_2 appears to be removed prior to atmospheric oxidation. Due to the rapid rate of SO_2 removal (half-life of less than 12 hours) regional transport of SO_2 also appears to be less important than ground deposition. Based on these rough analyses, the scope of the remainder of this section is restricted to ground deposition processes.

GROUND DEPOSITION

Ground deposition processes have gained recognition in the last 3 to 5 years as a major sink for SO_2 and other atmospheric pollutants. Studies done in England in regions dominated by grass fields or wheat crops indicate that over 40 to 60 percent of emitted SO_2 could ultimately be removed in one of the ground deposition processes.[2,3,4] Recent work in the RAPS Study Region (Metropolitan St. Louis) has also indicated strong ground deposition.[5] While these processes have not been adequately studied it is logical to assume that differential ground deposition characteristics are responsible for some of the observed variation in air quality.

There are at least three distinct mass transfer mechanisms collectively described on ground deposition: 1) adsorption on soil and vegetation surfaces, 2) absorption in water droplets and layers, and 3) physiological uptake in vegetation.

1. The absorption process is a potentially reversible operation which is independent of plant characteristics. Absorption occurs only during periods when water droplets are on the leaf surfaces. This would primarliy be restricted to night times when the temperature drops below the dew point or during periods of rain. The absorption may or may not be irreversible depending on the fate of the droplets. It is logical to assume that pollutants absorbed in dew are released upon evaporation of the droplets. SO_2 collected during the night could be released during the warm late morning periods by this process.

   There are several factors which probably inhibit mass transfer to the water droplets. These droplets are present only during periods generally characterized by poor mixing, therefore, the diffusional boundary layer in the gas phase is at a maximum. Furthermore, the droplets are stagnant, therefore, capacity for SO_2 due to the rise in pH which results from absorption of SO_2 and CO_2. This pH limit is particularly important for rain droplets which have an initial pH of ≈ 3.5 to 4.5.

Absorption processes could be either controlled by the gas phase diffusional resistivity or the gas liquid solubility equilibrium. This depends primarily on the presence of alkaline dust on vegetation surfaces, the pH of rain in the region of interest and the quantity of water on the surfaces. If it is diffusion controlled, the rate is proportional to the 1½ power of the absolute temperature and to the ½ power of windspeed. The solubility step, however, is inversely proportional to temperature.

2. Adsorption is a reversible process which is directly proportional to the available surface area. Water vapor and CO_2 probably compete with SO_2 and other pollutants for the active sites, therefore, such gases influence the surface area relationship. Adsorption involves a gas phase diffusion step which limits the overall rate of mass transfer. According to standard mass transfer equations, this process should be directly proportional to the 1½ power of the absolute temperature and to the ½ power of the average windspeed. There should be only a weak dependence on the type of plants, however, the total available surface area would be important.

3. The physiological uptake process is inherently irreversible and highly dependent on plant characteristics. The process is more complicated with respect to mass transfer and is potentially more important than the surface adsorption/absorption mechanisms. The pollutant gases must traverse a diffusional boundary layer and a relatively stagnant stomatal cavity. Upon contact with a cell, the molecule is adsorbed on the cell wall prior to diffusion through the cell membrane. It is not presently known which diffusional mass transfer operation controls the uptake rate. The overall mechanism should be directly proportional to temperature within the normal tolerance limits of the plant.

   The opportunity for physiological uptake is limited to daylight

> hours during which photosyntheses is occurring. Certain defensive mechanisms could reduce this uptake since both SO_2 and CO_2 could eventually stimulate closure of the stomata. Water stress would also keep stomata closed to reduce transpiration losses. Physiological uptake is directly proportional to the leaf surface area. Unlike the surface removal processes, physiological uptake is favored during the warm months.

Physiological uptake introduces a seasonal and diurnal cycle which conceivably influences dose-rate. While this is not of direct interest in the present study which involves annual average data, the dose-rate of SO_2 could be two to four times lower in daytime periods in warm months compared with night time or winter periods.

ANALYSIS OF GROUND DEPOSITION PARAMETERS

Ground deposition processes conceivably exert a strong influence of annual average SO_2 concentrations. Therefore, an attempt has been made to incorporate ground deposition parameters into the statistical model used to predict SO_2 exposure in those counties lacking adequate monitoring equipment. The mass transfer analyses discussed earlier indicate that the most useful parameters include the following:

- vegetation surface area
- vegetation type
- temperature to 1½ or 2 power
- windspeed to ½ power

The latter two factors introduce a seasonal and diurnal cycle not of direct interest to the present study. The vegetation factor, however, should be useful in approximating the influence of ground deposition processes.

A limited analysis has been performed using the Statistical Analysis System (SAS) computational package in order to evaluate the impact of the vegetation parameters on the predictive capabilities of the ambient SO_2 emmission inventory models. The scope was limited to North Carolina counties due to the availability of land use/vegetation distribution data.

These data included a county by county tabulation of acreage devoted to crop land, pastures, forests and "other" uses in 1958 and 1967. The similarities of the 1958 and 1967 data sets indicated that land use is relatively stable and therefore, probably representative of 1975 conditions. The SO_2 ambient concentration data for 1975 was obtained from the SAROAD data files. Several sites reporting extreme SO_2 levels were not representative of general conditions in North Carolina. Furthermore, only sites located in residential areas were considered. The final data set and descriptive statistics are presented in Table C-8. It is apparent that there is a very small range in reported SO_2 levels compared with most industrial regions. This will complicate subsequent analyses.

Analyses were done using both Spearman Correlation Tests and Multiple Regression Models. The former were used as a screening procedure to indicate possible linear or non-linear relationships. The multiple regression models generally included various emission parameters and land use parameters. Two sets of weighting factors were used to approximate the expected influence of vegetation surface area. These factors, shown in Table C-9, represent minimum and maximum conditions.

TABLE C-8-a. COUNTY EMISSION DATA AND LAND USE CHARACTERISTICS FOR NORTH CAROLINA

County ID	SO_2 emissions, T/yr			SO_2 µg/m³	Land use, acres			
	Total	Point	Area		Crop land	Pastures	Other	Forest
80	348	32	316	6.04	36146	22884	4395	94363
480	22598	21008	1590	5.02	36000	41985	9685	251800
500	1008	115	893	7.41	30166	13656	19567	199110
520	2972	1778	1194	6.01	60948	28524	8387	99487
540	995	237	758	5.82	25971	22565	10854	175150
660	1486	144	1342	6.54	65547	20540	16616	123675
720	23230	22717	513	6.04	54010	29813	9916	343600
780	520	402	118	5.18	36312	2200	7231	64657
840	1043	48	995	5.15	101213	35932	8425	127500
940	3525	2928	597	5.18	71972	4772	18448	287057
960	2885	1113	1772	7.46	93068	12566	12816	224528
1140	1023	665	368	5.07	166617	9534	10574	325820
1300	1306	772	534	7.59	136907	13178	10237	152700
1480	5509	2485	3024	6.72	69234	20489	21000	115037
1560	93756	91856	1900	5.05	37944	18465	11273	111636
1780	6999	3129	3870	15.46	101666	30235	17025	192300
1840	1224	764	460	6.25	117719	11617	12546	232400
1860	10448	9812	636	5.26	19641	40544	8367	141402
1940	261	14	247	5.03	56647	3015	10031	157727
2060	1857	746	1111	7.71	108155	63247	9526	166372
2120	1042	216	826	5.62	211024	11500	16500	249200
2280	552	159	393	7.60	30700	7057	4100	111292
2360	567	143	424	5.57	64999	17558	8149	91411
2480	2760	1793	967	5.02	13169	10399	7396	177516
2580	3467	222	3245	9.59	61507	24000	16155	156100
2680	569	101	468	6.35	51478	11278	13871	351561
2880	38902	36717	2185	6.44	6330	906	10032	67880
2980	1179	297	882	5.11	58391	1750	23362	300716
3000	1804	1282	522	6.27	49025	25116	10828	154500
3080	598	319	279	5.08	54089	3000	4889	78228
3180	1254	548	706	5.35	158692	13622	22049	191637
3260	1507	364	1143	5.43	111759	21218	17893	296562
3320	514	59	455	6.34	49056	12591	10204	217554
3380	6318	5458	860	6.17	224569	17957	11762	317300
3420	18455	17394	1061	5.59	83382	15217	15222	233898
3460	26252	25008	1244	8.23	93202	35938	12801	150420
3500	13563	12307	1256	4.99	64288	18000	12521	248121
3600	913	617	296	6.84	50529	7293	6091	99477
3900	2039	1460	579	6.17	82260	35035	7948	113252
3980	244	14	230	5.04	7128	4975	2150	84203
4080	726	102	624	5.97	136684	50960	15837	185200
4120	797	303	494	6.57	44118	5955	6929	85578
4160	3332	1151	2181	8.66	107941	26277	24679	322521
4280	20315	19447	868	5.25	146664	9415	21623	159382
4360	956	261	695	4.99	107642	2637	5219	110000
4500	603	49	554	4.99	14884	22109	4192	120705

TABLE C-8-b. DESCRIPTIVE STATISTICS FOR MODEL

Variable	Number of Observations	Mean	Minimum	Maximum	Standard deviation
Ambient SO_2 ($\mu g/m^3$)	46	6.30	4.99	15.46	1.763
Total SO_2 emmissions (10^3T/yr)	46	7.22	0.24	93.76	15.56
Point source SO_2 emissions (10^3T/yr)	46	6.23	0.01	91.86	15.33
Area source SO_2 emmissions (10^3T/yr)	46	0.99	0.12	3.87	0.81
Crop land (10^3 acres)	46	76.29	6.33	224.57	50.63
Pastures (10^3 acres)	46	18.64	0.91	63.25	13.70
"Other" land (10^3 acres)	46	11.94	2.15	24.68	5.57
Forests (10^3 acres)	46	179.58	64.66	351.56	81.77

TABLE C-9. VEGETATION SURFACE AREA WEIGHTING FACTORS

Land use category	Weighting factors	
	Set A	Set B
Pasture	1	1
Crop land	2	5
Other	3	25
Forest	4	100

The results of the statistical tests are presented in Tables C-10 and C-11. Correlations with a level of confidence greater than 90 percent are marked with an asterisk. It is apparent that the area source SO_2 emmissions are the dominant influence on ambient SO_2 levels. This observation is consistent with most emission impact studies.

The land use parameters did not improve the predictive capability of models using area source SO_2 emissions data. Note that all Spearman coefficients of land use parameters were equal to or lower than the comparable model based solely on area source emissions. Similar results are apparent in the linear and multiple regression models. A multiple regression model using appropriate powers of the area source emissions yields a modest increase in the correlation coefficient (0.44 improved to 0.71). However, the slight improvement does not warrant the considerable effort involved in compiling the land use (vegetation surface area) parameter.

While the consideration of vegetation characteristics should improve the SO_2 emission ambient SO_2 models, the techniques evaluated in this section were unsuccessful. It is felt that this could be due to the small range of ambient SO_2 concentrations or a too simplistic approach to the ground deposition process.

TABLE C-10. SPEARMAN CORRELATIONS: ANNUAL AVERAGE SO_2 CONCENTRATION AND VARIOUS EMISSION AND LAND USE PARAMETERS

Parameter	Spearman coefficient	Prob > F
Total SO_2 emmissions	0.131	0.61
Point source SO_2 emissions	0.064	0.67
Area source SO_2 emissions	0.253	0.09
Total SO_2 emissions/county area	0.096	0.53
Area SO_2 emissions/county area	0.149	0.67
Crop land	0.160	0.28
Pasture land	0.259	0.08
"Other" land	0.269	0.07
Forest	0.016	0.91
Area source emissions/Set A[a]	0.252	0.09
Area source emissions/Set B	0.233	0.11
Area source + 10% point source emissions/Set A	0.187	0.21
Area source + 10% point source emissions/Set B	0.184	0.22

[a] Set A and Set B denote the two sets of weighting factors for land use categories as described in Table 1-2.

TABLE C-11. MULTIPLE REGRESSION MODELS:

ANNUAL AVERAGE SO_2 CONCENTRATION AS DEPENDENT VARIABLE

Independent variables	Correlation coefficient (R^2)	Prob > F
Total SO_2 emissions	0.004	0.67
Point source SO_2 emissions	0.010	0.50
Area source SO_2 emissions	0.439	0.0001
Crop land	0.0090	0.53
Pasture land	0.0610	0.10
"Other" land	0.0582	0.11
Forests	0.0004	0.88
Total emissions/area	0.0019	0.76
Area source emissions/area	0.1200	0.02
(Area source emissions)$^{\frac{1}{2}}$	0.339	0.0001[a]
(Area source emissions)^2	0.581	0.0001[a]
(Area source emissions)^3	0.652	0.0001[a]
Log (area source emissions)	0.234	0.0007[a]
Area source - Total	0.717	0.0001
Area sources	-	0.0001
(Area sources)^2	-	0.0001
(Area sources)$^{\frac{1}{2}}$	-	0.0151
(Area sources)^3	-	0.4365
Log (area sources)	-	0.0257

REFERENCES

1. Richards, J. R. and Richard Gerstle. Stationary Source Control Aspects of Ambient Sulfates: A Data Base Assessment. Office of Research and Development, U.S. Environmental Protection Agency, Research Triangle Park, N.C. 157 pages. February 1976.

2. Garland, J. A. Deposition of Gaseous Sulphur Dioxide to the Ground. In: Atmospheric Environmental. Great Britain. 8:75-79, July 1973.

3. Shepherd, J. G. Measurements of the Direct Deposition of Sulphur Dioxide Onto Grass and Water by the Profile Method. In: Atmospheric Environmental. Great Britain. 8:69-74, July 1973.

4. Garland, J. A., W. S. Clough, and D. Fowler. Deposition of Sulphur Dioxide on Grass. In: Nature. 242:256-257, March 23, 1973.

5. Wilson, W. Personal Communication, February 1975.

6. North Carolina Conservation Needs Inventory. N.C. Department of Agriculture. December 1971.